

# Selected Public Databases and Software Tools with Relevance to the ESIGNET Project

C. Schmidt, T. Hinze, T. Lenser, P. Dittrich

Bio Systems Analysis Group  
Friedrich Schiller University Jena

[nyota@minet.uni-jena.de](mailto:nyota@minet.uni-jena.de)

December 13, 2006

keywords: cell signaling, pathway databases, protein databases,  
manual, visualization tools

## PREFACE

*This report aims to explain the use of selected databases that are useful information resources for the ESIGNET project (Evolving Cell Signaling Networks in Silico).*

*The overall goal of this project is to study the computational properties of cell signaling networks (CSNs) by evolving them using methods from evolutionary computation, and to re-apply this understanding in developing new ways to model and predict real CSNs. CSNs are of fundamental importance for our understanding of organismic processes and for medical research (understanding of diseases, development of effective drugs). A general and theoretical understanding of CSNs is currently missing: ESIGNET aims to fill this gap. For further information we refer to <http://www.esignet.net/>.*

*Almost all of the considered databases have many features and options and provide a lot of diverse information. We tried to find and to explain only those features of each database that are useful to know. With this knowledge it should be easy to use and understand other features. In particular, we have avoided to describe all the menus that could easily be discovered and understand. This will keep the report as clear as possible. We recommend to play around with a particular database while reading about it in this document.*

*Each chapter covers one database or software tool. There is a reference at the end of each chapter. Please note that we have list all the literature that we used and read for writing this report. It is not meant to be a recommended reading.*

*It is common that several different databases are integrated with cross-links into one comprehensive database. In such cases we only described the comprehensive database since it is the interface to the integrated databases. For retrieving the relevant information the user only needs to know about using this interface.*

# Contents

<b>CLASSIFICATION OF SIGNALING EVENTS</b>	<b>1</b>
<b>1 STCDB</b>	<b>1</b>
1.1 Browsing the classification system . . . . .	3
1.2 Searching for a certain signal transduction . . . . .	3
1.3 Overview of all STCDB signal transductions . . . . .	4
1.4 Background . . . . .	4
References . . . . .	4
<b>PROTEIN &amp; SIGNALING PATHWAY INFORMATION</b>	<b>4</b>
<b>2 UniProtKB</b>	<b>5</b>
2.1 Annotated information . . . . .	7
2.2 Querying UniProt . . . . .	7
2.3 Information of the entry's page . . . . .	7
References . . . . .	11
<b>3 SPAD</b>	<b>12</b>
<b>STRUCTURE AND ARCHITECTURE OF PROTEINS</b>	<b>16</b>
<b>4 PDB</b>	<b>16</b>
4.1 Searching for a protein structure & its properties . . . . .	18
4.2 Structure summary page . . . . .	20
4.3 Searching for binding-sites . . . . .	20
4.4 Searching for ligands . . . . .	21
4.5 Information about protein domains . . . . .	22
4.6 'Browse Database' – another kind of search . . . . .	23
References . . . . .	25
<b>5 InterPro</b>	<b>26</b>
5.1 Searching for a signature . . . . .	28
5.2 Explanation of an InterPro entry's structure . . . . .	28
5.3 Graphical representation of signatures . . . . .	29
References . . . . .	31

<b>INTERACTIONS</b>	<b>31</b>
<b>6 Ligand Depot</b>	<b>32</b>
6.1 Performing searches . . . . .	32
6.1.1 How to search for ligands by chemical name or formula . . . . .	32
6.1.2 How to search for PDB ligands by 3-letter ID . . . . .	33
6.1.3 Finding a PDB ligand by structure or substructure . . . . .	33
6.1.4 Browsing other sites containing ligand informations . . . . .	34
References . . . . .	34
<b>7 DIP</b>	<b>35</b>
7.1 Composition of DIP . . . . .	37
7.2 Data access and exchange . . . . .	37
7.3 Searching DIP . . . . .	37
7.4 Interaction graph . . . . .	40
7.5 Other search methods . . . . .	40
References . . . . .	42
<b>BINDING-SITES</b>	<b>42</b>
<b>8 CASTp</b>	<b>43</b>
8.1 About the input . . . . .	45
8.2 About the results . . . . .	45
8.3 Using CASTp . . . . .	45
References . . . . .	49
<b>9 GRASS &amp; Columbia Picture Gallery</b>	<b>50</b>
9.1 GRASS . . . . .	50
9.2 Columbia Picture Gallery . . . . .	51
<b>10 PASS</b>	<b>52</b>
References . . . . .	52
<b>APPENDIX</b>	<b>53</b>
<b>11 Tools for 3D-visualization of molecules</b>	<b>53</b>
11.1 PyMOL . . . . .	53
11.2 RasMol . . . . .	53
<b>12 Definitions</b>	<b>55</b>
Accession Number . . . . .	55
Binding Site . . . . .	55
Residue . . . . .	55
Protein Domain . . . . .	55
Protein Signature . . . . .	56
Asymetric Unit . . . . .	58

Biological Unit . . . . .	60
Gene Ontology (GO) terms . . . . .	62
Protein pocket . . . . .	63
Protein cavity . . . . .	64
References . . . . .	64



# Chapter 1

## STCDB

### **SIGNAL TRANSDUCTION CLASSIFICATION DATABASE**

<http://bibiserv.techfak.uni-bielefeld.de/stcdb/welcome.html>

With the widespread use of modern techniques in various subfields of biology, more and more cellular data are being accumulated, which has led to a proliferation of information and terminology. STCDB is based primarily on a proposed classification of signal transduction and it describes each type of characterized signal transduction for which a unique ST number has been provided and thus brings order into a classification recommendation. A systematic classification scheme is given for the various types of signal transduction and related reactions. Such systems for organizing and categorizing functions of bioprocesses (like the nomenclature of enzymes) are an important first step toward acquiring understanding of cellular processes. Moreover, consistent nomenclature is indispensable for communication and literature search.

With the help of signal transduction numbers it is possible to easily detect whether two signaling networks are the same or how similar they function (affiliation to the same class/category of function). No struggling with confusing names and synonyms! Another advantage of ST numbers is being able to make signaling pathway alignments which graphically show differences between networks. Pathways are represented as sequences of ST numbers and can be aligned with the software tool "PathAligner". It is also available from the University of Bielefeld.

The main source for the data in the STCDB comes from the CSNDB. A minor part of the data has been extracted from TransPath and BioCarta as well as from literature. Note that the CSNDB seems to not exist anymore.





## 1.1 Browsing the classification system

First, click on 'ST\_Classification' (1). Now you can get an overview about the classification system by browsing from the general categories to the more specific subcategories of signaling. When you have reached the most specific level of classification you might also find links to "BioCarta", "PubMed" and "CSNDB". Not every link is available for all cases. The link to CSNDB doesn't work. BioCarta has a short description of the network as well as a representation of it. With PubMed one can find literature and publications about the network.

**STCDB - Classification**

Browse in the ST Number tree (B):

- 1. Extracellular signal reception events
  - 1.1. Physical stimulation of signals (optical signals)
  - 1.2. Binding with hormones
    - 1.2.1. Hypothalamic peptide hormones
    - 1.2.2. Pituitary peptide hormones
    - 1.2.3. CGRP family
    - 1.2.4. Retinoic acid family
      - 1.2.4.1. 9-cis retinoic acid -> retinoic acid receptor
      - 1.2.4.2. 9-cis retinoic acid -> retinoic X receptor
      - 1.2.4.3. all-trans retinoic acid -> retinoic acid receptor
      - 1.2.4.4. retinoic acid -> retinoic acid receptor
    - 1.2.5. Other ungrouped peptide hormones
    - 1.2.6. Lipid hormones
  - 1.3. Binding with non-GF cytokines
  - 1.4. Binding with growth factors
  - 1.5. Binding with neuronal receptors
  - 1.6. Binding with Other ligands
- 2. Plasma membrane transduction events (plasma membrane -> plasma membrane)
- 3. Plasma membrane to cytoplasm transduction events (plasma membrane -> cytoplasm)
- 4. Intracellular signal transduction (cytoplasm -> cytoplasm)
- 5. Cytoplasm to nucleoplasm transduction events (cytoplasm -> nucleoplasm)
- 6. Nucleoplasm to nucleoplasm transduction events (nucleoplasm -> nucleoplasm)

[Home](#) | [Classification](#) | [Search](#) | [Network](#) | [Submission](#) | [Application](#) | [Contact](#)

Figure 1.2: STCDB: How browsing in the classification looks like.

## 1.2 Searching for a certain signal transduction

Click on 'ST\_Search' (2). Search either by ST number or by keyword. Searching by ST number is like browsing but a little bit faster since you can get the result directly. Searching by keyword is looking for a signal transduction that goes from molecule1 to molecule2 (like a ordinary chemical reaction). The keyword search is case-insensitive but the orthography is important: if a word is spelled wrong the search fails! Note that '\*' acts as a wildcard. For example,  $X \rightarrow *$  is searching the database for all transductions from X to any other molecule. Another use of the wildcard is completing a string: "abc\*" means all signal molecules which start with "abc".

### 1.3 Overview of all STCDB signal transductions

At 'ST\_Network' (3) one can find all networks covered by STCDB as an image. One can select either the whole network (consisting of many small networks) or networks of selected reactions. It is possible to zoom in regions of the image.

### 1.4 Background

#### About ST numbers and the classification system

A four-digit ST number  $d_1.d_2.d_3.d_4$  denotes a specific signal transduction where  $d_1$  is the location of transduction,  $d_2$  is the type of interaction,  $d_3$  describes the signal molecule's nature and  $d_4$  is a unique ID.

For example let us consider possible values and their meaning for  $d_1$ . Value  $d_1=1$  stands for extracellular signal reception events, whereas  $d_1=2$  stands for plasma membrane transduction events. There are six different values and meanings for  $d_1$ . Dependent on the value of  $d_1$  (the locations) there are different types of interactions:  $d_1 = 1.d_2 = 2$  is extracellular binding with hormones, whereas  $d_1 = 2.d_2 = 2$  is plasma membrane ion channel transduction. A full listing can be found on the web page.

### References

- 1 M. Chen, S. Lin, R. Hofstaedt:  
*STCDB: Signal Transduction Classification Database*;  
Nucleic Acids Research, 2004, Vol.32, D456-D458

## Chapter 2

# UniProtKB

### UNIVERSAL PROTEIN KNOWLEDGEBASE

<http://www.expasy.uniprot.org/>

UniProt is a single, centralized, authoritative resource for protein sequences and functional information. It is created by combining Swiss-Prot, TrEMBL and PIR. This makes it the world's most comprehensive resource on protein information. Swiss-Prot was recognized as the gold standard of protein annotation, with extensive cross-references, literature citations, and computational analyses provided by expert curators. Recognizing that sequence data were being generated at a pace exceeding Swiss-Prot's ability to keep up, TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was created to provide automated annotations for those proteins not in Swiss-Prot. The UniProt databases consist of three database layers: The UniProt Archive (UNIPARC) provides a stable, comprehensive sequence collection without redundant sequences by storing the complete body of publicly available protein sequence data. The UniProt Reference Clusters (UNIREF) databases provide non-redundant reference data collections based on the UniProt knowledgebase. UniRef90 and UniRef50 are built from UniRef100 to provide sequence collections to perform faster homology searches. All records from all source organisms with mutual sequence identity of >90% or >50%, respectively, are merged into a single record that links to the corresponding UniProt Knowledgebase records. The UniProt Knowledgebase (UNIPROTKB) is the central database of protein sequences with accurate, consistent, and rich sequence and functional annotation. The UniProt knowledgebase consists of two parts: a section containing fully manually annotated records resulting from literature information extraction and curator-evaluated computational analysis, and a section with computationally analysed records awaiting full manual annotation. For the sake of continuity and name recognition, the two sections are referred to as 'Swiss-Prot' and 'TrEMBL'.

This report focuses on UniProtKB since it seems to be most useful for the ESIGNET project.

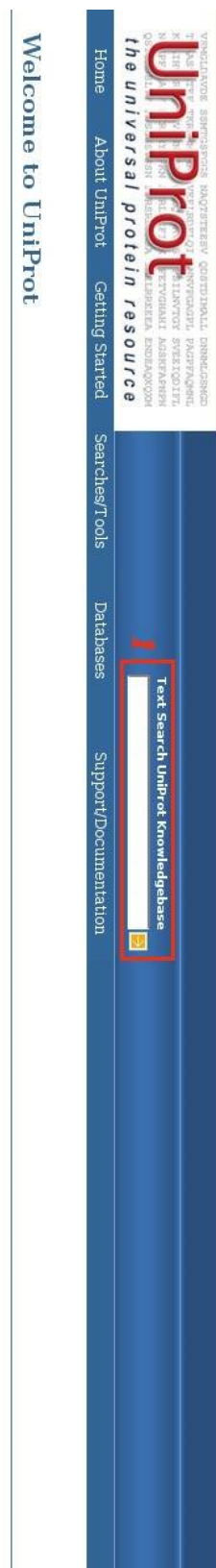


Figure 2.1: UniProt: Homepage.

## 2.1 Annotated information

In UniProtKB, annotation consists of the description of items like: function(s) of the protein, enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulation mechanisms), biologically relevant domains and sites, post-translational modification, molecular weight determined by mass spectrometry, tissue-specific expression of the protein, interactions, similarities to other proteins & diseases associated with deficiencies or abnormalities of the protein. This annotation is found in the comment lines, feature table and keyword lines.

## 2.2 Querying UniProt

The sequences and information in UniProt are accessible via text search, BLAST similarity search, and FTP. The most efficient and user-friendly way to browse the UniProt databases is via the UniProt web site. The web site provides database query mechanisms, user support and communication, file download capabilities, and links to related resources.

A keyword search in UniProtKB can easily be performed using the search bar (1). It seems that all information available on the results page are valid keywords. Hence one can search by PDB ID, protein name, entry name, accession number, species, author, pathway, ... If more than one database entry is found it should be selected either from 'Swiss-Prot' or 'TrEMBL' section.

## 2.3 Information of the entry's page

Blue items are links that start a new search with the item's name as a keyword. (Only sometimes it is a link to further details, like a text file opening in a browser window.) On top of the page is another search interface (2). Here, one can start a new query without having to go back to the UniProt homepage. On top of the page is also a row of links (3). These links refer to the information fields each entry is composed of. They are useful for navigating through the information provided.

UniProtKB is not really suitable for finding information about ligands, domains and binding-sites. But it provides a list of publications and many cross-references that might be a precious source of pursuing information.

[EXPASY Home page](#)
[Site Map](#)
[Search EXPASY](#)
[Contact us](#)
[Swiss-Prot](#)

Search

Swiss-Prot/TREMBL (full text)

▼

for

1iee

Go

Clear

## Search in UniProt Knowledgebase (Swiss-Prot and TREMBL) for: 1iee

UniProtKB/Swiss-Prot Release 50.4 of 25-Jul-2006: 230133 entries

UniProtKB/TREMBL Release 33.4 of 25-Jul-2006: 3051654 entries

---

Enter search terms:

1iee

Prefix and append wildcard '\*' to words.

Search in: ☐ Swiss-Prot ☐ TREMBL

new search

clear

---

By default, this search engine searches for complete words only. If you did not find what you expected, and would try to do a substring match, you should perform a new search and select 'prefix and append wildcard to words'.

---

- Number of documents found in UniProtKB/Swiss-Prot<sub>(1)</sub> and UniProtKB/TREMBL<sub>(0)</sub>: **1**
- For more directed searches, you can use the Advanced search form, or the Sequence Retrieval System SRS

---

**Search in UniProtKB/Swiss-Prot: There are matches to 1 out of 230133 entries**

LYSC\_CHICK (P00698)  
 Lysozyme C precursor (EC 3.2.1.17) (L4-beta-N-acetylmuramidase C) (Allergen Gal d 4) (Gal d IV). {GENE: Name=LYZ} - Gallus gallus (Chicken)

---

**Search in UniProtKB/TREMBL: There are matches to 0 out of 3051654 entries**

[EXPASY Home page](#)
[Site Map](#)
[Search EXPASY](#)
[Contact us](#)
[Swiss-Prot](#)

Figure 2.2: UniProt: After submission of a query all matches are shown.



EXPASY Home page
Site Map
Search EXPASY
Contact us
Swiss-Prot

2
Search
Swiss-Prot/TrEMBL
for
live
Go
Clear



Printer-friendly view  
Submit update  
Quick Blastp search  
Entry history

# UniProtKB/Swiss-Prot entry P00698

3
[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

*Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.*

<b>Entry information</b>	
Entry name	LYSC_CHICK
Primary accession number	P00698
Secondary accession number	Q90884
Integrated into Swiss-Prot on	July 21, 1986
Sequence was last modified on	July 21, 1986 (Sequence version 1)
Annotations were last modified on	July 25, 2006 (Entry version 78)
<b>Name and origin of the protein</b>	
Protein name	Lysozyme C [Precursor]
Synonyms	EC 3.2.1.17 1,4-beta-N-acetylmuramidase C Allergen Gal d 4 Gal d IV
Gene name	Name: LYZ
From	Gallus gallus (Chicken) [TaxID: 9031]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Phasianidae; Phasianinae; Gallus.
<b>References</b>	
[1] NUCLEOTIDE SEQUENCE, AND SEQUENCE REVISION TO 121. PubMed=6934509 [NCBI, EXPASY, EBI, Israel, Japan] Jung A., Sippl A.E., Grez M., Schutz G.; "Exons encode functional and structural units of chicken lysozyme."	
<b>Comments</b>	
<ul style="list-style-type: none"> <li>• <b>FUNCTION:</b> Lysozymes have primarily a bacteriolytic function; those in tissue and enhance the activity of immunogens.</li> <li>• <b>CATALYTIC ACTIVITY:</b> Hydrolysis of 1,4-beta-linkages between N-acetylmur between N-acetyl-D-glucosamine residues in chitodextrins.</li> <li>• <b>SUBUNIT:</b> Monomer.</li> <li>• <b>TISSUE SPECIFICITY:</b> In the egg white and polymorphonuclear Leukocytes.</li> <li>• <b>ALLERGEN:</b> Causes an allergic reaction in human.</li> <li>• <b>MISCELLANEOUS:</b> Lysozyme C is capable of both hydrolysis and transglyco peptide-substituted and unsubstituted peptidoglycan, and slowly on chitin olig</li> <li>• <b>SIMILARITY:</b> Belongs to the glycosyl hydrolase 22 family [view classification]</li> <li>• <b>WEB RESOURCE:</b> NAME = Worthington enzyme manual; URL = "http://www.worthington.com"</li> </ul>	

Figure 2.3: UniProt: Entry page.

## Some useful information fields and cross-references

The field 'Comments' is particularly useful. It shows an overview of an entry protein's biochemistry that is not provided in many other databases. The link 'Comments/Web Resource' is not available for every entry but it provides even more information about chemical properties.

For a number of proteins there is no knowledge about their interaction-networks. Hence the cross-reference 'Protein-protein interaction databases/DIP' isn't always working. "DIP" is covered more in detail in the Interactions section of this report.

'Other/ProtoNet' provides automatic hierarchical classification of protein sequences. Anyway, this is not so important for us. What might be interesting for us is a different representation of motifs and domains. Click on the link 'Get motifs and domains of protein' below the sequence representation of the respective UniProtKB entry. The GO terms might also be interesting.

'Features': The 'Feature Table' lines provide a precise but simple means for the annotation of the sequence data. The table describes regions or sites of interest in the sequence. In general the feature table lists posttranslational modifications, binding sites, enzyme active sites, local secondary structure or other characteristics reported in the cited references.

Motifs and domains of protein P-39667		1	10	20	30	40	50	60	70	80	90	100
PFAM	[1]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
PRINTS	[12]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
PROSITE	[1]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
SMART	[1]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
ACT_SITE	[2]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
BINDING	[1]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
CHAIN	[1]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
DISULFID	[8]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
HELIX	[7]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
SIGNAL	[1]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
STRAND	[9]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										
TURN	[13]	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQATNRNTDGGTDYGLQINSRWWCNDGRTPGSRNLCNIPCSALLSSDI										

Figure 2.4: ProtoNet: Another representation of motifs & domains.



Key	From	To	Length	Description
INIT_MET	0	0		
CHAIN	1	174	174	Ferritin light chain.
DOMAIN	6	155	150	Ferritin-like diiron.
REGION	53	60	8	Catalytic site for iron oxidation.
METAL	53	53		Iron ( <i>Potential</i> ).
METAL	56	56		Iron ( <i>Potential</i> ).
METAL	57	57		Iron ( <i>Potential</i> ).
METAL	60	60		Iron ( <i>Potential</i> ).
METAL	63	63		Iron ( <i>Potential</i> ).
MOD_RES	1	1		N-acetylserine.
CONFLICT	93	93		P -> L (in Ref. 2 and 3).
CONFLICT	135	137		DEE -> NEQ (in Ref. 3).
STRAND	4	4	1	
STRAND	7	7	1	
HELIX	10	36	27	
TURN	37	38	2	

Figure 2.5: UniProt: A part of the 'Feature Table' for Ferritin light chain (horse).

## References

- 1 R. Apweiler, A. Bairoch et al:  
*UniProt: the Universal Protein knowledgebase*;  
Nucleic Acids Research, 2004, Vol.32, D115-D119
- 2 A. Bairoch, R. Apweiler et al:  
*The Universal Protein Resource (UniProt)*;  
Nucleic Acids Research, 2005, Vol.33, D154-D159

## Chapter 3

# SPAD

### Signaling Pathway Database

<http://www.grt.kyushu-u.ac.jp/spad/>

SPAD is an integrated database for genetic information and signal transduction systems. SPAD is divided into four categories based on extracellular signal molecules (Growth factor, Cytokine, and Hormone) and stress, that initiate the intracellular signaling pathway. SPAD is compiled in order to describe information on interaction between protein and protein, protein and DNA as well as information on sequences of DNA and proteins. Currently, SPAD is both, under development and incomplete. It was last updated on Oct 13, 1998!

## Signaling Pathway Database

---

**The Signaling Pathway Database (SPAD) is an integrated database for genetic information and signal transduction systems.**

There are multiple signal transduction pathways: cascade of information from plasma membrane to nucleus in response to an extracellular stimulus in living organisms. Extracellular signal molecule binds specific intracellular receptor, and initiates the signaling pathway. Now, there is a large amount of information about the signaling pathway which controls the gene expression and cellular proliferation. We have developed an integrated database SPAD to understand the overview of signaling transduction. SPAD is divided to four categories based on extracellular signal molecules (Growth factor, Cytokine, and Hormone) and stress, that initiate the intracellular signaling pathway. SPAD is compiled in order to describe information on interaction between protein and protein, protein and DNA as well as information on sequences of DNA and proteins.

There are two methods for retrieving this database. Please select one of the two items.

### • Extracellular Signal Molecules



This WWW service "SPAD" is still under development.

---

mail to: [hirakawa@gnt.kyushu-u.ac.jp](mailto:hirakawa@gnt.kyushu-u.ac.jp)

[Molecular Gene Technics](#)

[Hakozaki Higashi-ku,](#)

[Fukuoka, 812-8581, Japan](#)

[Graduate School of Genetic Resources Technology](#)

[Kyushu University](#)

*Last Update Oct 13, 1998*

Figure 3.1: SPAD: Homepage.

## How to work with the database

There should be two methods for retrieving the database. In fact there is only one: Click on 'Extracellular Signal Molecules'. Four categories are listed. Every blue link goes to an interactive signaling pathway map. By clicking on the proteins one gets a summary for either receptor proteins or mediator proteins. It is not possible to click on the reactions connecting the proteins on the map. 'Journal Information' links are obsolete and down.

SPAD is classified into the four categories based on extracellular signal molecule that initiate the intracellular signaling pathway.

### 1. Growth Factor

[1-1. PDGF \(Platelet-Derived Growth Factor\)](#)

[1-2. EGF \(Epidermal Growth Factor\)](#)

[1-3. TGF-beta \(Transforming Growth Factor-beta\)](#)

[1-4. EPO \(Erythropoietin\)](#)

[1-5. NGF \(Nerve Growth Factor\)](#)

[1-6. IGF \(Insulin-like Growth Factor\)](#)

[1-7. TPO \(Thrombopoietin\)](#)

### 2. Cytokine

[2-1. IFN-alpha \(Interferon-alpha\)](#)

Figure 3.2: SPAD: Select by extracellular signal molecule.


# MAPK

<b>LOCUS</b>	HSU07620	2372 bp	mRNA	PRI	06-APR-1994
<b>DEFINITION</b>	Human MAP kinase mRNA, complete cds.				
<b>ACCESSION</b>	U07620				
<b>KEYWORDS</b>	.				
<b>REFERENCE</b>	1 (bases 1 to 2372)				
<b>AUTHORS</b>	Mohit,A.A., Martin,J.H. and Miller,C.A.				
<b>TITLE</b>	p49 3F12 kinase: A novel MAP kinase expressed in a subset of neurons in the human nervous system				
<b>JOURNAL</b>	Unpublished				
<b>STANDARD</b>	full automatic				
<b>REFERENCE</b>	2 (bases 1 to 2372)				
<b>AUTHORS</b>	Mohit,A.A.				
<b>TITLE</b>	Direct Submission				
<b>JOURNAL</b>	Submitted (09-MAR-1994) Abdi A. Mohit, Pathology, University of Southern California, School of Medicine, 2011 Zonal Ave, Los Angeles, CA 90033, USA				
<b>STANDARD</b>	full automatic				
<b>FEATURES</b>	CDS				
	224..1492				
	/note="NCBI gi: 468151"				
	/map="21q"				
	/product="MAP kinase"				
	/codon_start=1				
	/translation="MSLHFLYYCSEPTLDVKIAFCQGFQKQVDSYIAKHYNMSSKV				
	DNQFYSVEVGDSFTVLKRYQLKPIGSGAQGVCAAYDAVLDRNVAIKKLSRPFQNO				
	THAKRAYRELVLKMCVNHKNIISLLNVFTPKTLEEFQDVYLVMELMDANLCQVIQME				
	LDHERMSYLLYQMLCGIKHLHSAGIHRDLKPSNIVVKSDCTLKILDFGLARTAGTSF				
	MMTPYVTRYRAPEVILGMGYKENVDIWSVGCIMGEMVRHKILFPGRDYIDQWNKVI				
	EQLGTPCPEFMKKLQPTVRNYVENRPKYAGLTFFPKLFPDSLFPADSEHNKLKASQARD				
	LLSKMLVIDPAKRISVDDALQHPYINWYDPAEVEAPPPQIYDKQLDEREHTIEEWKE				
	LIYKEVMNSEKTKNGVVGQSPSAQVQQ"				
<b>MEDLINE ID</b>	no MEDLINE ID				
<b>OTHERS</b>	<a href="#">mapk and Human</a>				


Figure 3.3: SPAD: Results page for proteins in a signaling cascade.

Plasma  
Membrane  
receptor


Erythropoietin receptor


 **Synonym:**

- Erythropoietin receptor

 **Structure/Function**

The murine receptor is a protein of 507 amino acids with a single membrane-spanning domain. The cytoplasmic domain has a length of 236 amino acids. A point mutation at codon 129 of the murine Epo receptor gene results in constitutive activation. Mice expressing the aberrant receptor develop erythrocytosis and splenomegaly. Clonal growth factor-independent, proerythroblast cell lines that express Epo receptor have been isolated from the spleen of these animals.

 **[Search GenBank\(through Genome Net\)](#)**

 **[Sequence Information](#)**


 **[Journal Information](#)**

Figure 3.4: SPAD: Results page for proteins that are receptors.

## Chapter 4

# PDB

### PROTEIN DATA BANK

<http://www.rcsb.org/pdb/Welcome.do>

The Protein Data Bank is a central repository for 3-D structural data of proteins and nucleic acids. When the PDB was originally founded it contained just 7 protein structures. Since then it has undergone an approximate exponential growth in the number of structures. The coordinates of a structure are saved in a .pdb text file. Each PDB file's name is a unique ID, e.g. "1aew".





## 4.1 Searching for a protein structure & its properties

Figure 4.1 shows the search panel (1). One can search by PDB ID or by keyword which could be a protein's name, e.g. "ferritin light chain" or the name of a structure's author, e.g. "hilgenfeld". Querying the PDB by keyword can result in multiple hits (up to several hundreds). Different structures can refer to the same molecule but differ in resolution, origin (species) or by the experiment they were obtained from. There are structures with similar names, too.

However, if the search ended ambiguously one can either browse the hits to find the structure of interest or one can refine the search. The panel on the left provides this option. Here one can restrict the results by adding details to the search. 'Evaluate Subquery' checks the actual number of hits generated after adding a new detail. Tabular reports are another way to handle ambiguous search results. The option 'Tabulate' shows the results in a clear tabular fashion by presenting only certain information, e.g. showing a list of PDB entries and their ligands ('Tabulate/Summary Reports/Ligands'). So assume that you got too many hits and you know the ligand of the molecule that you are interested in. Then the tabular report is a way to find your structure more easily. The option 'Sort Results' might also be useful since it can sort the hits according to properties like resolution or date of publication.

It is also possible to start querying the PDB with an 'Advanced Search' (2). Another way of accessing search options is the 'Search' tab below the PDB logo (3). There one can access the 'Advanced Search' (2), too. It lets you add subqueries to the search like author, domain classification, ligand, disease, or EC number. Another useful feature of the 'Search' tab is 'Search Database/Ligands' (4). This loads an interactive drawing tool to sketch in the structure of a ligand. After drawing the tool searches the PDB for entries including this ligand. But searching for ligands in the PDB is more convenient by using "Ligand Depot". This web service will be covered later in its own chapter.



Contact Us | Help | Print Page

PDB ID or keyword

Author

SEARCH

Advanced Search

Home

Search

Structure

Queries

7

IAEW

Download Files

FASTA Sequence

Display Files

Display Molecule

Structural Reports

Structure Analysis

Geometry

Summaries and Analysis

8

PDBSum

MMDB

OCA

CSU Contacts

MSD (wwPDB partner)

DSSP Secondary Structure

Protein Motions

CASTp

Columbia Surface Picture Gallery

EDS

Classification

EBI Quaternary Structure

SCOP

CATH

3Dee

HSSP

FATCAT

VAST

DALI

Structure Summary

Biology & Chemistry

Materials & Methods

Sequence Details

Geometry

6

1AIEW

Title

L-CHAIN HORSE APOFERRITIN

Authors

Hempstead, P.D., Yewdall, S.J., Lawson, D.M., Harrison, P.M., Artymluk, P.J.

Primary Citation

Hempstead, P.D., Yewdall, S.J., Fernie, A.R., Lawson, D.M., Artymluk, P.J., Rice, D.W., Ford, G.C., Harrison, P.M., Comparison of the three-dimensional structures of recombinant human H and horse L ferritins at high resolution. *J.Mol.Biol.* v268 pp.424-448, 1997

History

Deposition 1997-02-26 Release 1997-09-04

Experimental Method

Type X-RAY DIFFRACTION Data [EDS]

Parameters

Resolution [Å]	R-value	R-Free	Space Group
1.95	0.192 (obs.)	n/a	F 4 3 2

Unit Cell

Length [Å]	a	b	c
Angles [°]	alpha	beta	gamma
	184.00	184.00	184.00
	90.00	90.00	90.00

Molecular Description

Polymer: 1 Molecule: FERRITIN Fragment: L-CHAIN Chains: -

Functional Class

Iron Storage

Source

Polymer: 1 Scientific Name: **Equus caballus** Common Name: **Horse** Expression system: **Equus caballus**

9 Chemical Component

Identifier	Name	Formula	Drug Similarity	Ligand Structure	Ligand Interaction
CD	CADMIUM ION	Cd <sup>2+</sup>	[View]	[View]	[View]

10 SCOP

Domain Info	Class	Fold	Superfamily	Family	Domain	Species
-------------	-------	------	-------------	--------	--------	---------

Images and Visualization

Biological Molecule

5

Display Options

KING

Jmol

Webmol

Protein Workshop

QuickPDB

All Images

Figure 4.2: PDB: Structure Summary.

## 4.2 Structure summary page

The result of the search for a macromolecule is the structure summary. It gives information about authors, the experiment, classifications, as well as molecular functions and properties. It is possible to visualize the structure with Java applets (5). Thin blue words are links that start a new PDB search with the word being a keyword. That makes it easy e.g. to get all known structures that share the same domain architecture, that have the same molecular function or that were determined by the same author.

More information about that found molecule can be obtained from the five tabs (6) below the search bar. Interesting are the 'Sequence Details': it provides a graphic of secondary structures mapped to the corresponding part of the sequence. In the top of that page there is a 'Domains' link. Following this link one can get the exact domain boundaries (from amino acid to amino acid). But in general, many of the information available can only be understood by experts. If you would like to get an example just take a look at 'Geometry'.

The panel on the left side (7) is important. Save a .pdb file with 'Download Files'. It may only contain the asymmetric unit of the structure. The functioning complete molecule can be saved with 'Biological Unit Coordinates'. To get more information about the concepts of the asymmetric unit and biological unit please read in the "Definitions" chapter at the end of this report.

Another interesting overview about the selected structure can also be found at the left panel (7): 'Structure Analysis/Summaries and Analysis/OCA'.

'Structure Analysis/Summaries and Analysis/MSD' gives also a summary about the selected structure but has a new good feature: 'Similarity'. It shows entries similar to the selected structure.

'Structure Analysis/Classification'/DALI': The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank. A multiple alignment of structural neighbours is mailed back to you. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences. If you want to know the structural neighbours of a protein already in the Protein Data Bank, you can find them in the FSSP database.

## 4.3 Searching for binding-sites

Unfortunately, there is only less information about binding-sites available from the PDB. Reasons for this situation could be difficulties in determining binding sites experimentally and computational and that binding-sites are of great value for companies that research for drugs. However, an overview about binding sites can be obtained from the structure summary of a certain protein. Go to 'Structure Analysis/Summaries and Analysis/PDBSum' (8) on the left panel. The picture 'Clefts' is a link to the cleft analysis. One can visualize

pockets and clefts on the surface of the protein. Those regions are potentially binding-sites. Visualization is done by the Jmol applet or with a script for RasMol which works well with PyMOL, too. To find out more about these two software tools read in the chapter "Tools for 3d-visualization of molecules".

There is also free software available that is capable of computing binding-sites as well as web services dealing with protein surface topography (see part "BINDING-SITES" for details). Note that the services "CASTp" and "Columbia Picture Gallery" can be accessed from within the PDB ('Structure Analysis/Summaries and Analysis').

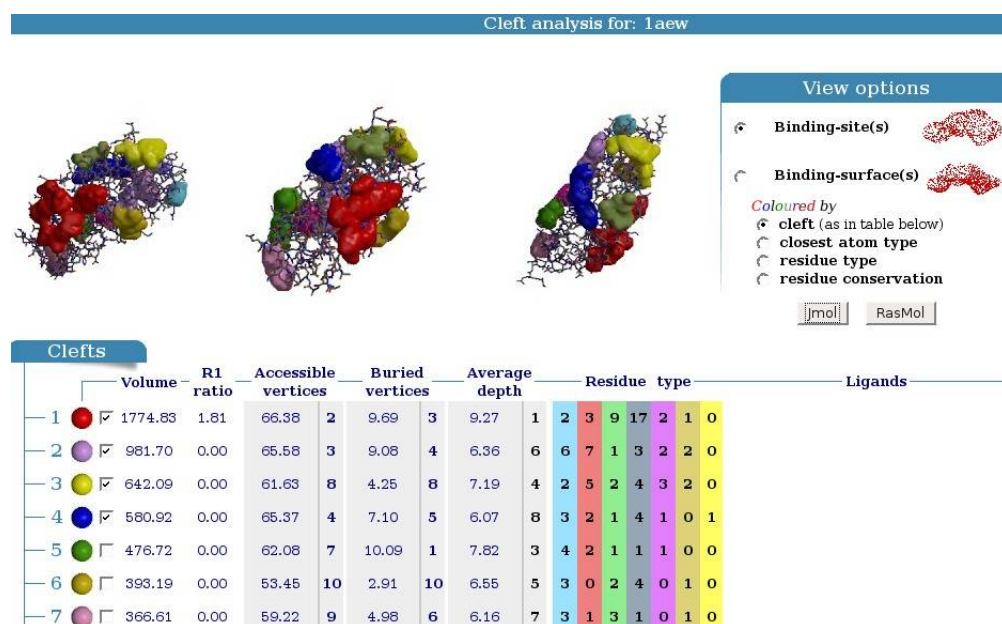


Figure 4.3: Cleft analysis with PDBSum.

## 4.4 Searching for ligands

One can do an advanced search for ligands as described above in the first section. Another way is going to the structure summary page of the protein of interest. Click 'Biology and Chemistry' (6). There one can find 'Ligands and Prosthetic Groups' which lists the name of the ligand and its chemical formula. A visualization of its structure is available, too. Note that it seems that the PDB entries are listing only those ligands that were bound to the protein during the experiment. On the structure summary page there is the entry 'Chemical Component' (9) which gives a list of ligands associated with the molecule. Note that it doesn't list the number of ligands.

Ligand Depot is a database that will find PDB entries containing a particular ligand. It is covered in its own chapter in part "INTERACTIONS".

There is a online service that analyses ligand-protein contacts. Unfortunately a software tool ("Chemscape") is needed that is not available for Linux. So visualization of such contacts is not possible. You can reach "LPC-Software" from the summary page's left panel (7) 'Structure Analysis/Summaries and Analysis/CSU Contacts'. But still this link could be useful since it has the number of ligands in a particular PDB entry. One can select single ligands for the analysis. The results cover data about a specific contact like residues that are making contact to the ligand, shortest distances, putative hydrogen bonds between ligand and protein, or contact area. It looks like all listed contacts types are non-covalent ones. Therefore, if the actual analysis doesn't match one of these contact types it might be a covalent contact.

Left panel (7) 'Structure Analysis/Summaries and Analysis/MSD/Ligand' provides information about ligands, too. There is a statistics about interactions of the ligand with different residues. The data is shown as a bar chart. Clicking on each bar reveals the total number of contacts with the respective amino acid. By clicking on that bar again one gets the PDB entries where the ligand interacts with the residue.

## 4.5 Information about protein domains

Go to the structure summary.

- In the left panel (7) click on 'Structural Reports'. The tab 'Sequence Details' (6) has a link called 'Domains' which provides a domain description and domain boundaries.
- Left panel (7) 'Structure Analysis/Classification'. Click on either 'SCOP', 'CATH' or '3Dee' for domain classification information. Moreover, SCOP provides a list of PDB entries sharing the same classification.
- At the bottom of the structure summary page there is the SCOP and CATH hierarchical classification of protein domains (10). This is the information that is available from both databases.

### About CATH & SCOP – Protein Structure Classification

Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. Thus classification is useful to clarify such similarities and to distinguish between groups of similar proteins.

The **CATH** database is a hierarchical classification of protein domains into sequence and structure based families and fold groups. In the lowest level of the hierarchy, sequences are clustered according to significant sequence similarity. At higher levels, domains are grouped according to whether they share significant sequence, structural and/or functional similarity. Fold groups are sharing similar architectures. These similarities in the arrangements of secondary structures are then merged regardless of their connectivity into

common architectures. At the top of the hierarchy, domains are clustered depending on their class, that is the percentage of  $\alpha$  helices or  $\beta$  strands (class1: mainly alpha, class2: mainly beta, class3: mixed alpha-beta, class4: domains which have low secondary structure content).

So there are four major levels in this hierarchy (from top to bottom): class, architecture, topology (fold family) and homologous superfamily.

Similar to CATH **SCOP** classifies proteins of known structure into families, superfamilies, folds and classes. The classification is on hierarchical levels that embody different levels of evolutionary and structural relationships between the domains.

In brief: CATH and SCOP are classifying proteins according to their domains, but they do not tell something about the function of protein domains.

## 4.6 'Browse Database' – another kind of search

The option (11) to browse the PDB can be found in the 'Search' tab (3). Figure 4.4 shows all categories that can be browsed. Each category is structured hierarchically (like a tree). Browsing through the hierarchy displays the structures belonging to each level. The lower (more specific) a level is the less structures belong to it.

Alternatively, the categories can be searched by keyword. Here of course, a PDB number is not a valid keyword! (Because the purpose of browsing is getting all PDB entries that belong to a certain hierarchical level.) A molecule name isn't a proper keyword either. It only works if this keyword is part of a item's name in the hierarchy, e.g. a metabolic pathway name.

If you browse the "Gene Ontology" (categories: 'Biological Process', 'Cellular Component' & 'Molecular Function'), please note that not all PDB IDs/chains have been mapped to GO terms.



## References

- 1** F. Pearl, C. Bennett, J. Bray et al:  
*The CATH Database: an extended protein family resource for structural and functional genomics;*  
Nucleic Acids Research, 2003, Vol.31, No.1
- 2** A. Murzin, S. Brenner, T. Hubbard, C. Chothia:  
*SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures;*  
J. Mol. Biol., 1995, 247, 536-540

## Chapter 5

# InterPro

<http://www.ebi.ac.uk/interpro/>

Secondary protein databases on functional sites and domains are vital resources for identifying distant relationships in novel sequences, and hence for predicting protein function and structure. InterPro is a comprehensive documentation resource for protein families, domains and functional sites. It combines a number of databases (referred to as member databases) that use different methodologies and a varying degree of biological information on well-characterised proteins to derive protein signatures. By uniting the member databases, InterPro capitalises on their individual strengths, producing a powerful integrated diagnostic tool. Currently, it includes PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF and SUPERFAMILY. Signatures are manually integrated into InterPro entries that are curated to provide biological and functional information. Each InterPro entry is described by one or more signatures, and corresponds to a biologically meaningful family, domain, repeat or site, e.g. post-translational modification. Entries are assigned a type to describe what they represent, which may be family, domain, repeat, PTM, active site or binding site. InterPro entries are annotated with a name, an abstract, mapping to Gene Ontology (GO) terms and links to specialized databases. InterPro groups all protein sequences matching related signatures into entries.

Protein signature databases have become vital tools for identifying distant relationships in novel sequences and hence are used for the classification of protein sequences and for inferring their function.



InterPro IPR000014 PAS				Click here for help!
<b>Matches</b> ?	Overview: Detailed: Table: Architectures	<a href="#">sorted by AC,</a> <a href="#">sorted by AC,</a> <a href="#">For all matching proteins,</a>	<a href="#">sorted by name,</a> <a href="#">sorted by name,</a> <a href="#">of known structure</a>	<a href="#">of known structure,</a> <a href="#">proteins with splice variants</a> <a href="#">of known structure</a> <a href="#">proteins with splice variants</a>
<b>Accession</b> ?	IPR000014 PAS Matches: 6786 proteins			
<b>Type</b> ?	Domain			
<b>Signatures</b> ?	Database	ID	Name	Proteins
	PROSITE profile	PS50112	PAS	5470
	SMART	SM00091	PAS	5385
	TIGRFAMs	TIGR00229	sensory_box	5208
<b>Found in</b> ?	IPR011785 Transcriptional regulator PpsR IPR012129 Phytochrome A/B/C/D/E IPR012130 Photoactive yellow-protein IPR012226 Oxygen sensor diguanylate cyclase/c-di-GMP phosphodiesterase IPR012704 Propionate catabolism, transcriptional regulator PPR IPR013655 PAS fold-3 IPR013656 PAS fold-4 IPR013767 PAS fold			
<b>Process</b> ?	GO:0007165 signal transduction			
<b>Function</b> ?	GO:0004871 signal transducer activity			
<b>Abstract</b> ?	<p>PAS domains are involved in many signalling proteins where they are used as a signal sensor domain. PAS domains appear in archaea, bacteria and eukaryotes. Several PAS-domain proteins are known to detect their signal by way of an associated cofactor. Haeme, flavin, and a 4-hydroxycinnamyl chromophore are used in different proteins. The PAS domain was named after three proteins that it occurs in:</p> <ul style="list-style-type: none"> <li>• Per- period circadian protein</li> <li>• Arrt- Ah receptor nuclear translocator protein</li> <li>• Stim- single-minded protein.</li> </ul> <p>PAS domains are often associated with PAC domains <a href="#">IPR001610</a> . It appears that these domains are directly linked, and that together they form the conserved 3D PAS fold. The division between the PAS and PAC domains is caused by major differences in sequences in the region connecting these two motifs [ 1 ]. In human PAS kinase, this region has been shown to be very flexible, and adopts different conformations depending on the bound ligand [ 2 ]. Probably the most surprising identification of a PAS domain was</p>			

Figure 5.1: InterPro: The structure of the entry page.

## 5.1 Searching for a signature

For example, consider the query "pas". PAS is both, a domain and a motif. The search finds PAS, PAC motif and PAS fold. Select one entry and get more informations about it as well as proteins containing the respective signature.

## 5.2 Explanation of an InterPro entry's structure

When selecting an InterPro entry from the search results one attains its page of information. (In other words, this page is the entry itself.) Each InterPro entry is composed of information fields like Header, Matches, Accession, Signatures, Relationships, Example proteins and so on. This report will not list and explain every information field, because the explanations are provided online by clicking on the field's name. Still, some fields are worth mentioning:

'Matches' gives a number of different views of the signature (InterPro entry) matches on all protein sequences containing the signature. The option 'Matches/Architectures' shows InterPro domain architectures, count of, example and architecture code. Domain architectures are displayed as a series of non-overlapping domains. For each InterPro entry, a graphical representation of unique domain architectures is provided and each kind of domain architecture is displayed with an example protein and total number of proteins, sharing this architecture, next to it. Clicking on the count of proteins retrieves all proteins sharing a common architecture.

'Accession' provides the number of proteins containing the signature.

Entries may be related to each other through two different relationships: The parent/child relationship is useful for indicating family/subfamily relations where the child (subfamily) is more specific than its parent. That's why a signature matching the child protein is always matching the parent, too. The second type of relationship is the contains/found in relation that indicates domain composition. Found in suggests that the signature/domain may be found in the listed proteins. These proteins on the other hand may do contain this signature. Parent/child relationships are used to describe a common ancestry between entries, whereas the contains/found in relationship generally refers to the presence of genetically mobile domains.

Useful information about the signature/domain can be obtained from 'Process', 'Function' and 'Abstract'. 'Database links/PROSITE doc' provides molecular biological information about the signature/domain.

### 5.3 Graphical representation of signatures

There is a compact view and a detailed view of the protein sequence and its matching signatures. In the compact view all signatures are shown in one row whereas the detailed view represents the protein sequence as a series of different lines for each protein signature hit. The protein sequence is represented as a scaled horizontal grey line, the protein match line, along which vertical lines are drawn at 10, 20, 50, 100, 200 or 500 amino acid intervals, depending on the length of the protein. The scale is shown to the left of the match graphics.

Coloured bars are displayed along the protein match line to indicate where in a protein matches were found among the InterPro entries. The bar is coloured according to which InterPro entry (e.g. PAS) matched that region of the protein. In addition to matches to InterPro entries, matches to curated structural data, CATH, SCOP, and PDB and to non-curated predicted structural elements defined by SWISS-MODEL and MODBASE are also displayed. The matches to these structural models have fixed colours with white striped lines. Moving the mouse over a coloured bar will show more information such as the residues corresponding to the position of the match on the protein.

See the key near the bottom of the page to identify which colours correspond to which InterPro entries or structural features. Clicking on the AC number of a protein takes you to its detailed view, which also shows the domain architecture view for this protein. This view represents the domain composition by oval shapes that contain the name and the number of iterations of the domain if greater than one. If there are more than 25 proteins on one page they are split into groups, using the sort order. The index (S-Swiss-Prot, T-TrEMBL) is shown on the left side of the page. Click on each section to view subsets of the selected proteins.

Note that splice variants of a protein are marked light-yellow, whereas ordinary proteins are marked light-blue.

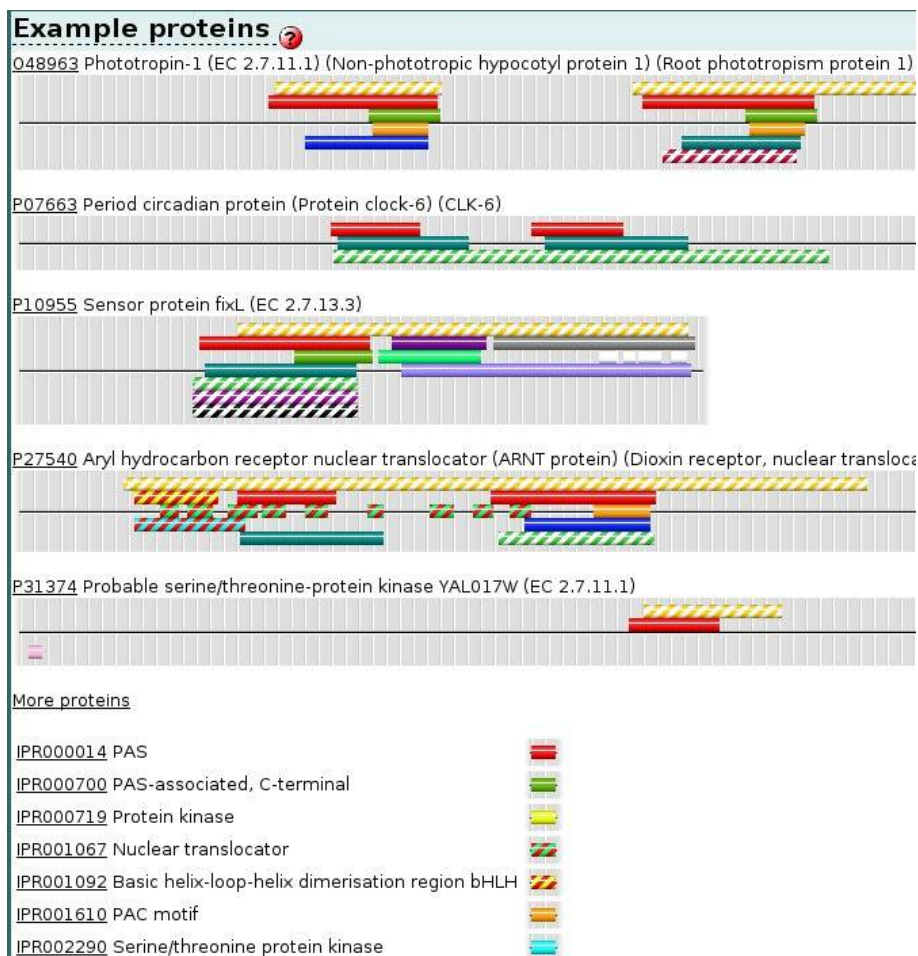


Figure 5.2: InterPro: Signatures of 'Example proteins'

Count Example Code	Architecture
478 007384 IDA13767,3661,3594	PAS fold His kinA N ATP bd ATPase
269 000327 IDA1092,13767,13655	HLH basic PAS fold PAS 3
252 P20169 IDA3660,13767,3661,3594	His kin HAMP PAS fold His kinA N ATP bd ATPase
224 028757 IDA3661,3594	His kinA N ATP bd ATPase
222 015055 IDA13655	PAS 3
210 052237 IDA13656,3661,3594	PAS 4 His kinA N ATP bd ATPase
191 024117 IDA13654,3018,13515,13767x2,3661,3594	PAS 2 GAF Phytochrome Cr PAS fold His kinA N ATP bd ATPase
135 067220 IDA13767,160,1633	PAS fold GGDEF EAL
133 031394 IDA13767	PAS fold
105 005250 IDA13767,3594	PAS fold ATP bd ATPase
102 01AU92 IDA13656,160,1633	PAS 4 GGDEF EAL

Figure 5.3: InterPro: Architecture view

## References

- 1     *InterPro User Manual*
- 2     N. Mulder, R. Apweiler et al:  
      *InterPro, progress and status in 2005*;  
      Nucleic Acids Research, 2005, Vol.33, D201-D205
- 3     N. Hulo, A. Bairoch et al:  
      *The PROSITE database*;  
      Nucleic Acids Research, 2006, Vol.34, D227-D230
- 4     C. Bru, E. Courcelle et al:  
      *The ProDom database of protein domain families: more emphasis on 3D*;  
      Nucleic Acids Research, 2005, Vol.33, D212-D215
- 5     I. Letunic, R. Copley et al:  
      *SMART 5: domains in the context of genomes and networks*;  
      Nucleic Acids Research, 2006, Vol.34, D257-D260

## Chapter 6

# Ligand Depot

<http://ligand-depot.rcsb.org/>

Ligand Depot is an integrated data resource for finding information about small molecules bound to proteins and nucleic acids. It focuses on providing chemical and structural information for small molecules found as part of the structures deposited in the Protein Data Bank. Ligand Depot accepts keyword-based queries and also provides a graphical interface for performing chemical sub-structure searches. A wide variety of web resources that contain information on small molecules may also be accessed through Ligand Depot. One can search for ligands in four different ways: by PDB chemical component ID, by name, by chemical structure, or by chemical formula.

### 6.1 Performing searches

The homepage of Ligand Depot offers two search fields. The upper field, is for ligand search in the PDB. The lower one is for accessing other web resources as described in 6.1.4.

#### 6.1.1 How to search for ligands by chemical name or formula

1. Select 'chemical name' or 'chemical formula' from the dropdown menu.
2. Select the option of matching the chemical name or formula exactly ('Equal') or partially ('Like').
3. Enter the chemical name or formula in the text box.  
Please note that searches are not case-sensitive. A range of atom numbers may be entered in the chemical formula (e.g. C34-36 H32-36 N4-5 O4-5 FE1), if desired.
4. Click on the Search button.

5. Click on the 3-letter PDB ID of a ligand of interest in order to obtain more information about that ligand.

If there is more information in other databases, a new window pops up in order to browse all of these databases or to access a single one manually.

### 6.1.2 How to search for PDB ligands by 3-letter ID

Ligands may be searched using their assigned 3-letter PDB ID. Please note that the assignment of PDB IDs is arbitrary and may not reflect the actual compound name or synonym in any way.

1. Select 'PDB component id' from the dropdown menu.
2. Enter a ligand ID consisting of 1 to 3 letters.
3. In the query results, clicking on a 3-letter PDB ID (e.g. CFF) will return a ligand report for the small molecule of interest.

### 6.1.3 Finding a PDB ligand by structure or substructure

A ligand may also be identified by performing a structural comparison between the ligand of interest and all of the small molecules present in the PDB. A graphical file containing a ligand's chemical structure may be uploaded into the drawing tool or else a molecule may be drawn from scratch. Ligand Depot performs the structural search by comparing the atoms and bonds of both the queried and target molecules. The results may include either the actual ligand being queried, or else larger ligands that contain the queried molecule as a substructure. Therefore, the structural comparison will return a list of PDB ligands that are structurally similar, either in part or in whole. Please note that the chemical substructure search can only be performed on ligands present in the PDB.

Graphical file formats that may be uploaded into the drawing tool include the mmCIF (macromolecular Crystallographic Information File) format and the MOL format.

Please note that the current drawing tool only works in Internet Explorer, Netscape and Mozilla (not Mozilla Firefox!).

1. Between the two search fields is the link 'Find a PDB ligand by structure or substructure'. Click on it.
2. Select the filetype from the dropdown menu.
3. 'Browse' and 'Load' your file.
4. After the ligand appears in the drawing tool, select Clean Up Sketch
5. Alternatively, instead of uploading a ligand file in CIF or MOL format, a ligand may be drawn from scratch using the drawing tool.
6. Modify the uploaded or sketched chemical structure as desired.

7. Click on the 'Add Hydrogens' button to search for the exact chemical structure.  
OR  
Click on the 'Remove Hydrogens' button to search for small molecules that contain the substructure of interest.
8. Select Clean Up Sketch again.
9. Select Search Substructures.
10. Click on the 3-letter PDB ID of a ligand of interest in order to obtain more information about that ligand.

The drawing tool's name is "MarvinSketch". A user manual can be found at <http://www.chemaxon.com/jchem/marvin/chemaxon/marvin/help/about-sketch.html> or <http://ligand-depot.rutgers.edu/marvin/chemaxon/marvin/help/sketch-index.html>.

#### 6.1.4 Browsing other sites containing ligand informations

Currently, Ligand Depot stores information from 70 small molecule sites. The different resources are organized into four categories including nomenclature sites, molecular visualization sites, commercial sites, and chemical databases. Selecting one of these categories returns a list of web resources with a brief description of what each one has to offer.

1. Select the desired site type from the dropdown box.
2. Click on the 'Browse' button.
3. Results are displayed in a new browser window.
4. Select a web resource of interest from the list of results.
5. The selected web site will be displayed in a second browser window and may then be searched for relevant information.

## References

- 1 Z. Feng, L. Chen et al:  
*Ligand Depot: a data warehouse for ligands bound to macromolecules*;  
Bioinformatics, 2004, Vol.20 no.13, pages 2153-2155



## Chapter 7

# DIP


### DATABASE OF INTERACTING PROTEINS

<http://dip.doe-mbi.ucla.edu>

The Database of Interacting Proteins aims to integrate the diverse body of experimental evidence on protein-protein interactions into a single, easily accessible online database. It provides a comprehensive and integrated tool for browsing and extracting information about protein interactions. By interact DIP means that two amino acid chains were experimentally identified to bind to each other. Because the reliability of experimental evidence varies widely, methods of quality assessment have been developed and utilized to identify the most reliable subset of the interactions. This core set can be used as a reference when evaluating the reliability of protein-protein interaction data sets, for development of prediction methods, as well as in the studies of the properties of protein interaction networks. The evaluation methods are implemented as publicly available services (<http://dip.doe-mbi.ucla.edu/dip/Services.cgi>) that can be used to evaluate the reliability of new experimental and predicted interactions.

DIP contains pairwise interactions between proteins and allows the visual representation and navigation of protein-interaction networks. The quality of a given interaction can be assessed visually by the thickness of the lines between two proteins and the selection of a specific method can be applied to show the results from only a given method. The DIP allows the integration of a diverse body of information onto a protein-interaction network, such as the predominance of certain domains or the different subcellular compartments in which a protein can be found.

This page serves also as an access point to a number of projects related to DIP, such as LiveDIP, The Database of Ligand-Receptor Partners (DLRP) and JDIP. Registration is required to gain access to most of the DIP features. Registration is free to the members of the academic community.



D
atabase of
I
nteracting
P
roteins

[SEARCH:TOP][LOGIN]

Help
News
Register
Statistics
Satellites
Services
Articles
Search
Links
Files
Jobs

The DIP™ database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data. Please, check the [reference](#) page to find articles describing the DIP database in greater detail.

**1**
This page serves also as an access point to a number of projects related to DIP, such as LiveDIP, The Database of Ligand-Receptor Partners ([DLRP](#)) and JDIP.

THE DIP DATABASE

DIP PAGES

NEWS

REGISTRATION/ACCOUNT

STATISTICS

SATELLITES

SERVICES

ARTICLES

Announcements about the most recent additions and changes to the database.

Registration and account maintenance. Registration is required to gain access to most of the DIP features. Registration is free to the members of the academic community. Trial accounts for the commercial users are also available. Please, consult [Terms of Use](#) for further details.

Detailed information about the current state of the database as well as some statistics on server usage.

DIP-related projects, such as [DLRP](#), [LiveDIP](#) and [JDIP](#).

DIP-derived services.

DIP in press. Both, papers published on DIP as well as a list of publications referring to DIP.

Figure 7.1: DIP: Homepage.

36

## 7.1 Composition of DIP

The DIP database is composed of nodes and edges. **DIP Nodes (proteins):** Each protein participating in a DIP interaction is identified by a unique identifier of the form <DIP:nnnN> and cross-references to, at least, one of the major protein databases - PIR, Swiss-Prot and/or Genbank. In addition, some basic information about each protein, such as name, function, subcellular localization and cross-references to other biological databases is stored locally (if available) in case the cross-referenced databases are not accessible.

**DIP Edges (interactions):** The information about each DIP interaction is identified by a unique identifier of the form <DIP:nnnE> that provides access to information such as the region involved in the interaction, the dissociation constant and the experimental methods used to identify and characterize the interaction.

## 7.2 Data access and exchange

The interactive, web-based interface allows users to query the database for a specific protein based on its name, annotation or species of origin. In case the protein of interest is not yet present in the database, it is also possible to perform sequence similarity (BLAST) and motif searches in order to identify closely related proteins. The pattern of interaction of these might provide insights into the potential but not yet identified interactions of the query protein.

In the batch mode, different subsets of the DIP database can be downloaded in a variety of formats ranging from the native XML-based XIN format to simple, tab-delimited text files that are ready to be imported into spreadsheet applications. The DIP data are also provided in the Molecular Interaction Format (MIF) developed under the auspices of the Human Proteome Organization (HUPO) Proteomics Standards Initiative.

## 7.3 Searching DIP

In order to start the exploration of the protein-protein interaction network the DIP database can be searched in a variety of ways to find the initial protein of interest. It is also possible to search for entire groups of proteins fulfilling certain criteria, such as sequence similarity to a given protein, specific function or cellular localization, the presence of a specified domain (e.g. InterPro, Pfam domains) or well-known sequence motif (e.g. Prosite motifs).

Click on (1) to start searching. Search types are 'Node', 'BLAST', 'Motif', 'Article' & 'PathBLAST'. Unfortunately the search is not robust. For example if you are looking for "glucokinase" (hexokinase gamma (HXKG)) you get only the entry for E.coli. But there is also glucokinase/yeast stored in the database. If one want to get interaction information from DIP and want to

search by protein name, InterPro AC, or UniProtKB entry name then one has to perform a 'Node Search'. It is a good idea to use UniProtKB entry names or other database's AC numbers.

1. Go to "Search/Node"  
Note that the field "Name/Description" accepts complex, logical expressions (OR, AND, NOT, brackets (), wildcard %)
2. Either one queries using 'Node Identifier' or 'Node Annotation'.  
If searching by protein name one has to use 'Node Identifier'. Otherwise no hits are generated by the search. Alternatively, one can type in entry names and AC numbers. Of course, this only works if the protein of interest has an entry in the respective databases. Using 'Node Annotation' turned out to be a bit unreliable.
3. One gets 'Node Search Results'.  
More information is available when clicking on the AC (2) below 'Node'.
4. A new window pops up containing two important links.  
'graph' displays the interaction graph and 'GO Function' displays a GO term description of the protein.


NODE SEARCH RESULTS					
DIP		Cross Reference			Protein Name/Description
Node	Links	PIR	SWISSPROT	GENBANK	
DIP:525N		JT0482	HXKG_YEAST	gi:83035	glucokinase
2					

Figure 7.2: DIP: Result of a Node Search.

DIP NODE						[graph]	
						[-----]	
DIP 525N		PIR	JT0482	SwissProt	HXKG_YEAST	RefSeq	NP_009890
Glk1 p	Name/Description						glucokinase
Organism	Saccharomyces cerevisiae						
Xrefs	GO Function						Help
G	HSSP: P05708	SGD: GLK1	GermOnline: 138876				MIPS: YCL040W
P	SGD: S0000546	PIR: JT0482	NCBI: X59720, M24077				
PRINTS: PR00475							PROSITE: PS00378
Pfam: PF00349							InterPro: IPR001312

Figure 7.3: DIP: Cross-references for a DIP entry.

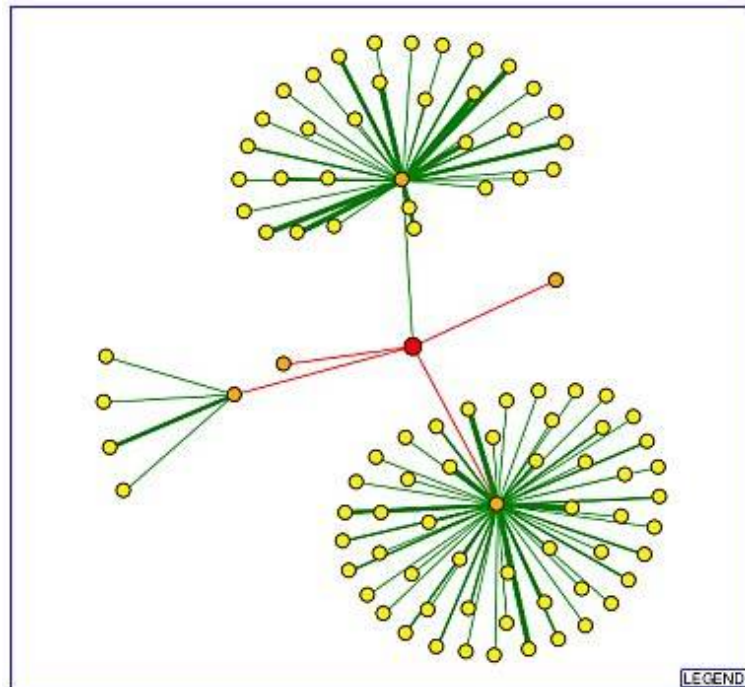


Figure 7.4: DIP: An example interaction graph.

DIP NODE						<a href="#">[graph]</a>
						<a href="#">[----]</a>
DIP 525N	PIR	<a href="#">JT0482</a>	SwissProt	<a href="#">HXKG_YEAST</a>	RefSeq	<a href="#">NP_009890</a>
Glk1p	Name/Description glucokinase					
Organism	Saccharomyces cerevisiae					
<a href="#">Xrefs</a>	GO Function					<a href="#">Help</a>
Process	carbohydrate metabolism ( <a href="#">GO:0005975</a> )					
Function	glucokinase activity ( <a href="#">GO:0004340</a> )					
Localizaton	cytosol ( <a href="#">GO:0005829</a> )					

Figure 7.5: DIP: GO functions for a DIP entry.

## 7.4 Interaction graph

The red node represents the queried protein. One can click on every node in the graph to get its information which is displayed in the same window as shown in figure 7.3. In order to maintain clearness of representation not all edges are drawn in the graph. (Nodes two edges away from root are displayed without their linking edges. Only edges from the first shell nodes are drawn.) The width of edges encodes the number of independent experiments identifying the interaction. Color encodes the reliability of the interaction evidence. Green is used to draw core interactions that were verified by one or more computational verification methods. The unverified results of the high-throughput interaction screens are drawn in red.

## 7.5 Other search methods

The first described kind of search is probably the most frequently performed search. Still there are other useful search methods. One can do a BLAST search. A given protein sequence is compared with all available protein sequences. All similar sequences are returned. It is also possible to search for a MOTIF. It is best to perform such a search in InterPro (e.g. keyword "HXKG\_YEAST") . Its information field 'Signatures' provides the ID of the respective PROSITE pattern. (And thus it suffices to only query InterPro instead of all its member databases!) This PROSITE ID in turn will be accepted by DIP which makes it easy to search for motifs. One can search with a 'Costum Pattern' as well. Note that searching this way will reveal glucokinase/yeast and hexokinase A & B! This corresponds to the InterPro 'Abstract' that says that there are three isozymes of hexokinase in yeast.

PATHBLAST does not belong to DIP but there is a cross-reference. It searches the protein-protein interaction network of the target organism to extract all protein interaction pathways that align with a pathway query. The query consists of protein ID's (e.g. UniProtKB entries) or protein sequences. The found networks are represented graphically.

Protein ID

A

ST11\_YEAST

and/or

Protein Sequence

MIKERPLNKASMIKSPVDEIYMEQTQTAEGLDIIIGDEKTNDI

B

STE7\_YEAST

and/or

MFQRKTLQRRNLKGLNLNLHPDVGNNGQLQEKTEHQQQSRII

C

KSS1\_YEAST

and/or

MARTITFDIPSQYKLVDLIGGAYGTVCSAIHKPSGIKVAIKI

+

Add a Protein

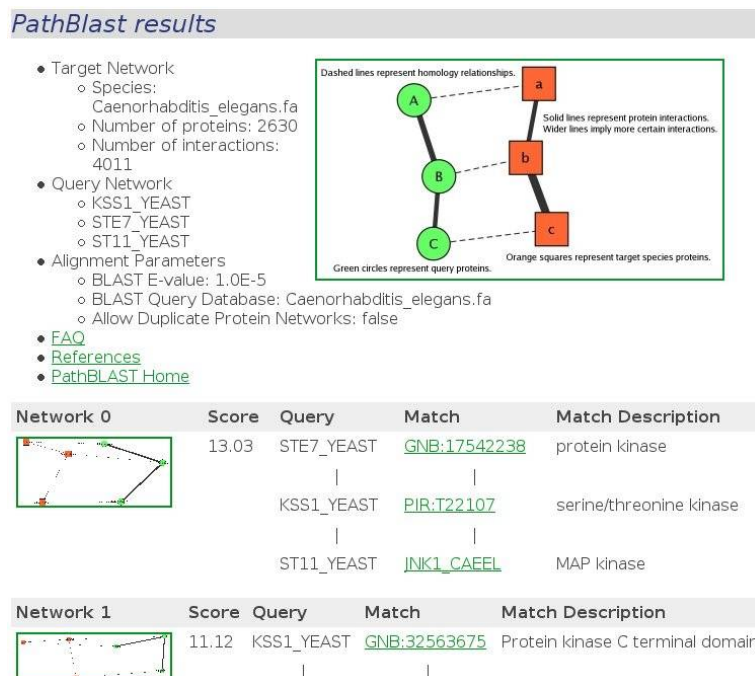
-

Remove a Protein

Please select the Target Organism Network:

Caenorhabditis elegans

Figure 7.6: PathBLAST: The query form.



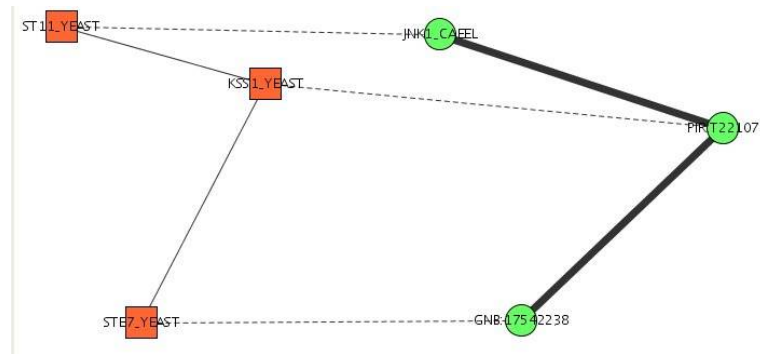


Figure 7.8: PathBLAST: The interaction network no. 0 shown in figure 7.7.

## References

- 1 L. Salwinski, C. Miller et al:  
*The Database of Interacting Proteins: 2004 update*;  
Nucleic Acids Research, 2004, Vol.32, D449-D451
- 2 I. Xenarios, L. Salwinski et al:  
*DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions*;  
Nucleic Acids Research, 2002, Vol.30, 303-305




## Chapter 8

# CASTp

### COMPUTED ATLAS OF SURFACE TOPOGRAPHY OF PROTEINS

<http://sts.bioengr.uic.edu/castp/>

Binding sites and active sites of proteins and DNAs are often associated with structural pockets and cavities. The CASTp server uses the weighted Delaunay triangulation and the alpha complex for shape measurements. It provides identification and measurements of surface accessible pockets as well as interior inaccessible cavities, for proteins and other molecules. It measures analytically the area and volume of each pocket and cavity, both in solvent accessible surface (Richards' surface) and molecular surface (Connolly's surface). It also measures the number of mouth openings, area of the openings, circumference of mouth lips, in both surfaces for each pocket. You can request calculation for a particular molecule. The results will be shown on the screen or emailed to you. The emailed results include measured parameters for pockets, cavities and mouth openings, as well as listing of wall atoms and mouth atoms for each pocket. In addition, a downloadable PyMOL plugin (Which at present seems not to work properly: There is a problem operating the CASTpyMOL plugin with newer versions of python.) will help you to visualize the pocket of your interest. CASTp allows access to information of computed pockets and voids for structures in the Protein Data Bank (PDB). Note that the results from CASTp differ a bit from the results from PDBSum.



# CASTp

## Computed Atlas of Surface Topography of proteins

[Home](#)
[pvSoar](#)
[Liang Lab](#)
[Bioengineering](#)
[UIC](#)

Please submit bug reports to [Joe Dundas](#). You can still [search the old CASTp server here](#).

- [Search](#)
- [Calculation Request](#)
- [CASTp Information](#)
  - About CASTp
  - Background
  - Input & output
  - Examples
  - Publications
- [PyMOL plugin](#)

### Search CASTp database

**Query:**  
[Query help](#)

Search

**Keyword:**

Search

#### Visualization options

**jmol - jmol** is a free, open source molecule viewer. It is cross-platform, running on Windows, Mac OS X, and Linux/Unix systems. It only requires that you have a java enabled browser. Works best in Internet Explorer.

**Chime Plugin** - Chime is a browser plugin that runs under Windows, MacOS, and IRIX. This is the best choice for Windows 95/98/NT/2000/XP or IRIX. Linux user can also use this plugin by running a virtual windows environment ([VMware](#)). Chime is available from [MDL's website](#).

**E-mail only** - This option allows you to get results back via email only. No visualizations will be provided.

E-mail results to:

### About CASTp

Binding sites and active sites of proteins and DNAs are often associated with structural pockets and cavities. castp server uses the weighted Delaunay triangulation and the alpha complex for shape

Figure 8.1: CASTp: Homepage.

## 8.1 About the input

You can either type the 4 letter PDB code of a protein structure if it is available from Brookhaven protein databank (in this case the CAST server will fetch that structure), or, you can upload a structure to the CAST server for calculation. The structure of the molecule to be uploaded must be in PDB format. Please take care to remove all nonpolar H atoms, else they will not be recognized and will be assigned a default radius of 1.8 angstrom, which may result in a misleading calculation. This is particularly relevant to NMR structures. Do not request from PDB or upload a structure that contains multiple conformers, such as those seen in NMR structure. CAST does not know which one to pick. Instead, upload a file containing a single structure of your interest, for example by editing the original file. All hetero atoms will be treated as ligand and will be automatically removed from calculation. This includes solvent water molecules.

## 8.2 About the results

After calculation you can receive an email with the following files for the calculated results: queried pdb file, listing of all the pockets for the queried file, measurements about each pocket and cavity, measurements about mouth openings of each pocket, listing of the atoms about the mouth openings & a listing of all the annotated residues.

## 8.3 Using CASTp

An intuitive graphic user interface allows querying of the CASTp server by typing the four letter PDB name of a protein structure, by keyword searching or by submitting their own molecular structure in the PDB format. When querying by keyword, a list of relevant PDB structures are returned as obtained by redirecting to RCSBs PDB query site. The requested structure can be visualized using the "Jmol" applet. In addition to the simple manipulations built in to the Jmol applet the user interface also allows selective highlighting of individual pockets ('Pocket Information', left panel). Summary information of measurement of individual pocket and void is conveniently displayed in a scrolling menu. Selection of a specific pocket from this menu also reveals the wall atoms comprising the pocket in a separate small window. By typing in the name and number of a residue, the user can easily identify the pocket or void that contains a particular residue.

Moreover there is panel on the right ('Annotated Sites') where one can find the amino acids belonging to a particular pocket, their number in the sequence and the function they are contributing to (like 'ACT\_SITE' – active site or 'BINDING' – the amino acid is important for binding a substrate). Those amino acids are represented in red. The Jmol applet below is a representation of the protein sequence. Amino acids are shown there in a color indicating the pocket they belong to.

Additional information completing the idea of binding-sites of a protein is provided by the "Columbia Surface Picture Gallery" that is accessible from the PDB structure summary page. Here one can find pictures of molecular surfaces colored by chemical and physical properties, e.g. 'distance to a ligand'.

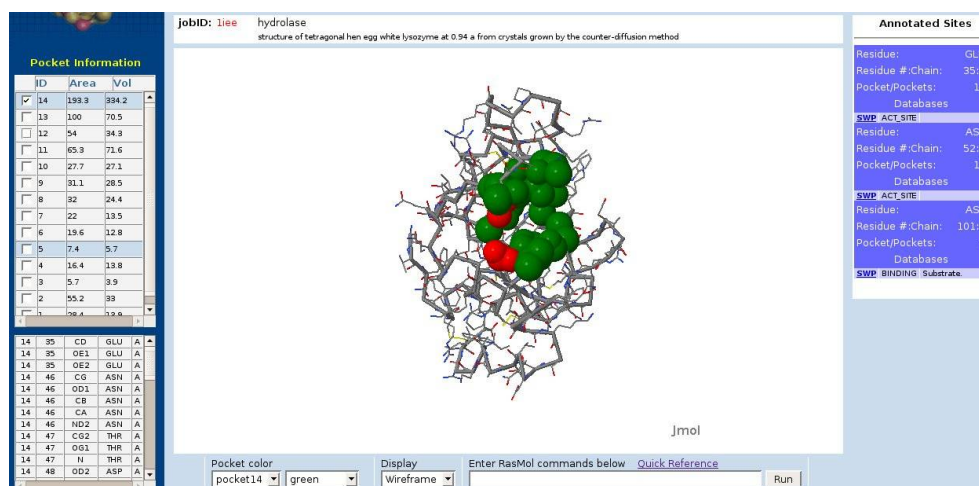


Figure 8.2: CASTp: Visualization of the computed pockets (1lee).

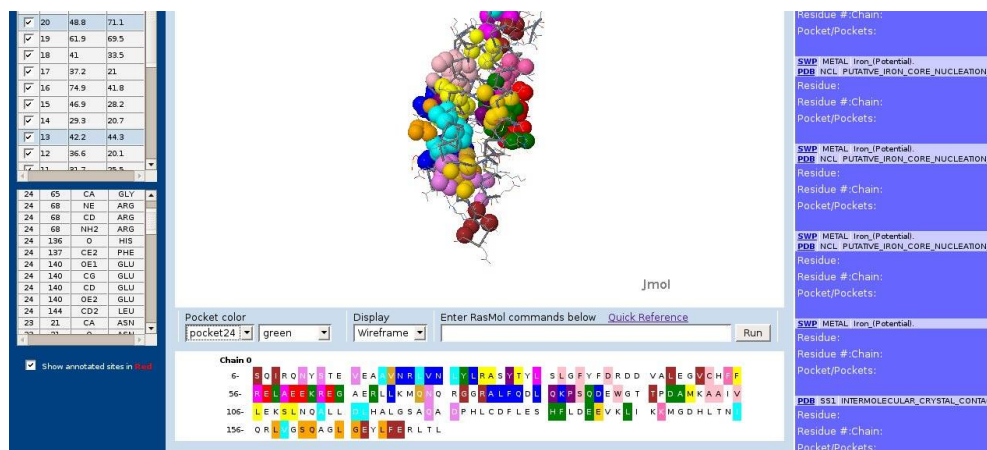


Figure 8.3: CASTp: Visualization of the computed pockets (1aew).

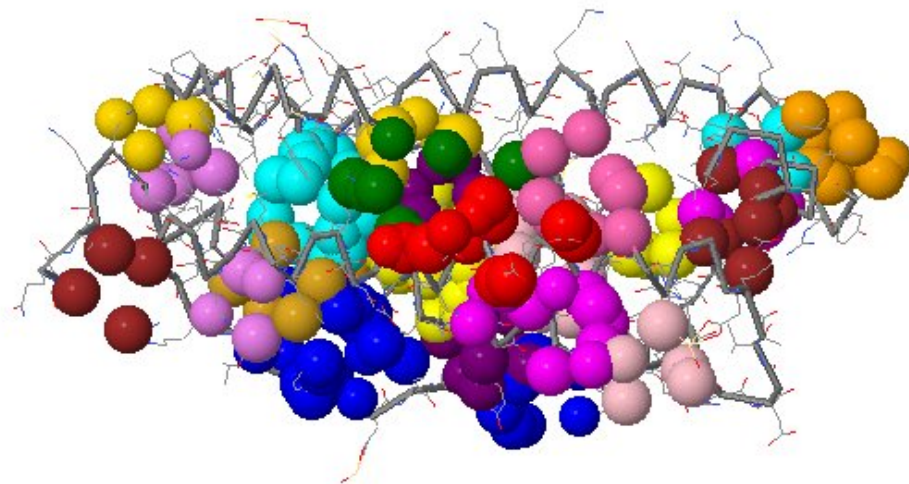


Figure 8.4: CASTp: Comparison of the results from CASTp and PDBSum (1aew in CASTp).

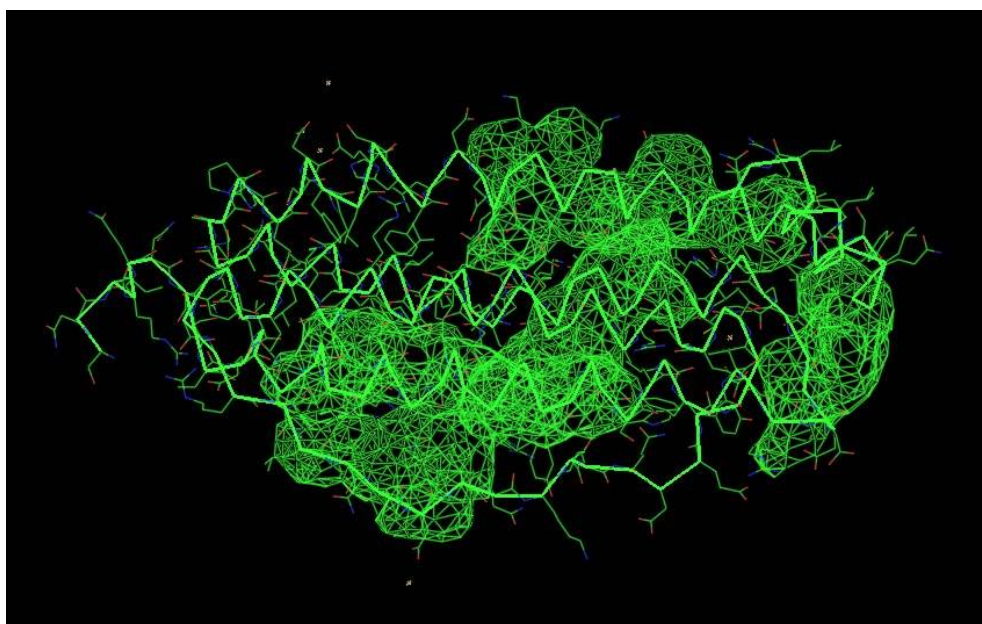


Figure 8.5: CASTp: Comparison of the results from CASTp and PDBSum (1aew in PDBSum).

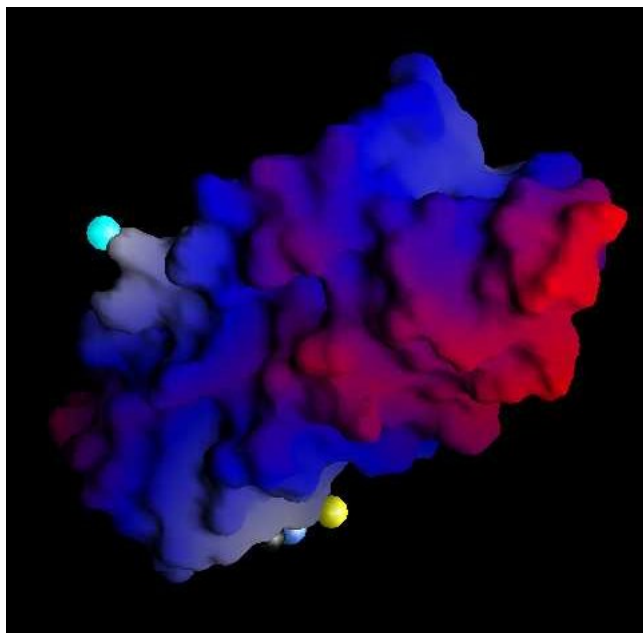


Figure 8.6: A Columbia Surface Picture of 1aew. The surface is colored by its distance to a ligand. The distance is increasing from white to blue to red. That means that the area underneath the dark yellow spheres in figure 8.4 is most distant from any ligand. There are actually no pockets in the protein surface. However figure 8.5 shows the PDBsum binding-sites in PyMOL. Here are some pockets displayed in this area.

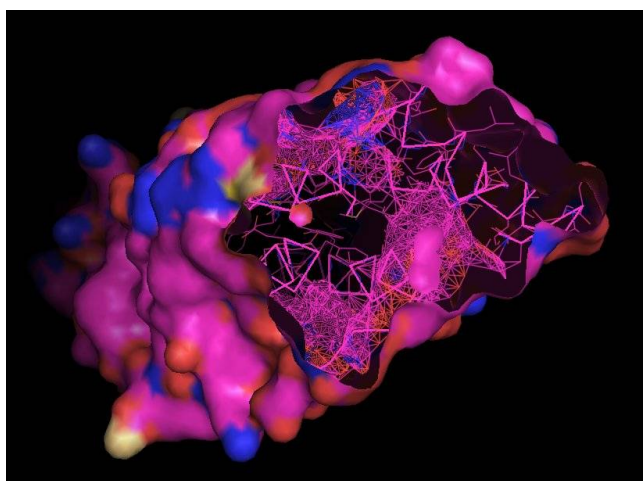


Figure 8.7: Binding-surface computation of 1aew. The binding-surface is represented by the thin lattice under the solvent generated surface (the smooth layer). The PDBsum result is displayed using PyMOL.

The figures show exemplarily that binding-sites cannot always be computed indisputable. That's why one should be careful with interpreting results. I would like to emphasize again that every computed pocket and cavity is not necessarily a binding-site. Even though binding-sites are known to be in pockets of a protein's surface.

Note that the picture from the Columbia Gallery is coloured by distance to a ligand. No pocket information is taken into account. This could be one reason why the results are not matching each other.

## References

- 1 T. Binkowski, S. Naghibzadeh, J. Liang:  
*CASTp: Computed Atlas of Surface Topography of proteins*;  
Nucleic Acids Research, 2003, Vol.31, 3352-3355



## Chapter 9

# GRASS & Columbia Picture Gallery

[http://honiglab.cpmc.columbia.edu/cgi-bin/GRASS/surfserv\\_enter.cgi](http://honiglab.cpmc.columbia.edu/cgi-bin/GRASS/surfserv_enter.cgi)

### 9.1 GRASS

GRASS (Graphical Representation and Analysis of Structures Server) exploits many of the features of the GRASP program and is designed to provide interactive molecular graphics and quantitative analysis tools with a simple interface.

GRASS is used in three steps: **First**, a macromolecular structure is selected for viewing and analysis. The user selects the molecules to be displayed by selecting one or more display styles for each. **Second**, the color-coding scheme for molecules is selected by choosing a molecular property to be calculated or fetched from a database. The molecule will be colored according to the molecular property chosen using a property value to RGB color correspondence. **Third**, one of three programs is selected to display the graphics: a VRML viewer, Chime or the GRASP molecular modeling program. Linux users have to use a VRML viewer. VRML is a general purpose three-dimensional scene description language.

I didn't have much time for testing. I used "Freewrl" and it didn't work properly. The browser plugin crashed several times and visualization was poor. However, this is not a serious drawback since one can easily access the precomputed pictures from Columbia.

It is planned to make a Linux version of the GRASP2 software available (for further details take a look at [http://wiki.c2b2.columbia.edu/honiglab\\_public/index.php/Software:GRASP2](http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:GRASP2)).



## 9.2 Columbia Picture Gallery

The picture gallery is accessible from the GRASS homepage mentioned above. It provides (static) pictures of the surface of a protein deposited in the PDB. The pictures are coloured by different chemical properties, for example like hydrophobicity, distance to a ligand or electrostatic potential. The pictures were generated using GRASS.

## Chapter 10

# PASS

<http://www.ccl.net/cca/software/UNIX/pass/overview.shtml>

### Fast Prediction of Protein Binding Pockets

PASS (Putative Active Sites with Spheres) is a simple computational tool that uses geometry to characterize regions of buried volume in proteins and to identify positions likely to represent binding sites based upon the size, shape, and burial extent of these volumes. The main utility of PASS lies in the fact that it can fastly analyze a moderate-size protein. PASS produces output in the form of standard PDB files, which are suitable for any modeling package, and provides a script file to simplify visualization in RasMol. PASS is freely available to all in unix executable form. So, installation amounts to simply decompressing and unarchiving the appropriate file.

PASS was shown to reliably predict the locations of known binding sites using a set of 20 apo-protein x-ray structures from the PDB, thereby establishing its utility as a front-end to fast docking and virtual screening. Furthermore, PASS provides the user a meaningful view of the buried volumes in a protein, suggests alternate binding sites, and simplifies detailed visualization of potential binding hot-spots.

As a modeling tool, PASS (i) rapidly identifies favorable regions of the protein surface, (ii) simplifies visualization of residues modulating binding in these regions, and (iii) provides a means of directly visualizing buried volume, which is often inferred indirectly from curvature in a surface representation.

### References

- 1 G. Brady, P. Stouten:  
*Fast Prediction and Visualization of Protein Binding Pockets with PASS*;  
DuPont Pharmaceuticals Company

## Chapter 11

# Tools for 3D-visualization of molecules

By now there are many software tools available that visualize molecules in 3D. Some of them are even capable of structure analysis. This report presents two useful and free molecular viewer. Both are running under Linux based operating systems and are easy to install. They are a good supplement to Java applet viewer that run inside a browser (and are provided by many databases).

### 11.1 PyMOL

PyMOL is being steadily improved. The source code is free for compilation (open source). The build can be used and installed for evaluation purposes. After expiration of the evaluation period one has to sponsor the project. PyMOL can display structures very beautiful and has analysis capability. With it one can even do real molecular modeling – building own structures or modifying structures. The handling is superior to all free molecular viewers that I have seen.

### 11.2 RasMol

Although RasMol runs unstable it seems to be much more known than PyMOL because it was one of the very first visualization tools. It can display the biological unit in a .pdb file (which PyMOL can't). Moreover there are many databases and software tools providing RasMol scripts. But often these work with PyMOL, too.

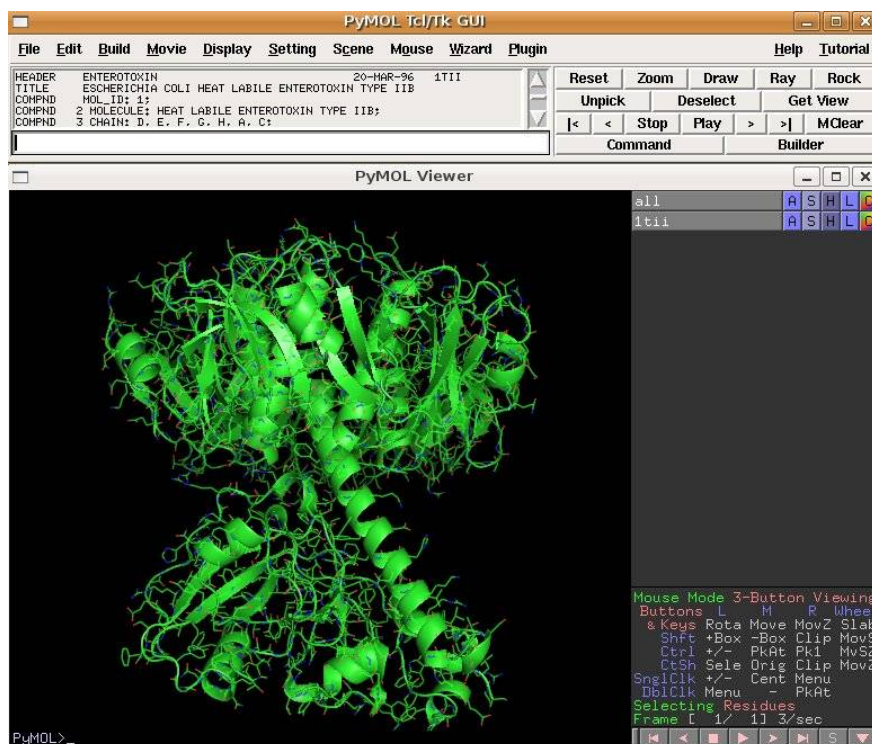


Figure 11.1: PyMOL displaying a pdb structure.

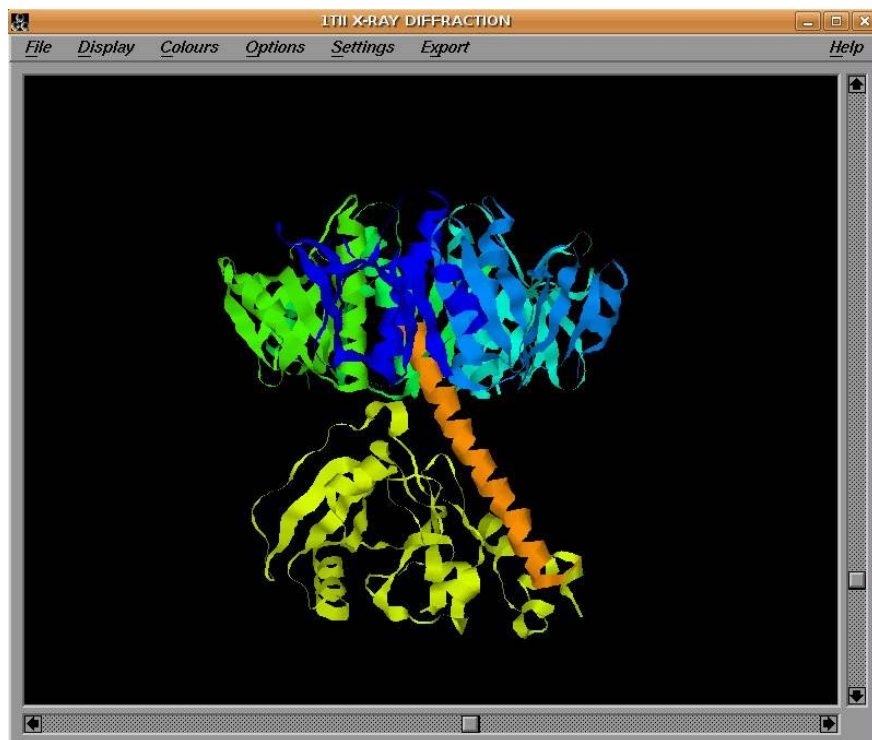


Figure 11.2: Rasmol displaying the same structure.

## Chapter 12

# Definitions

### Accession Number

The accession number (AC) provides a stable way of identifying database entries. (Every entry is mapped to one unique AC.)

### Binding Site

Protein performs its function through interaction with other molecules such as substrate, ligand, DNA and other domains of proteins. The three-dimensional structure of protein provides the necessary shape and physicochemical texture to facilitate these interactions. Sites of activity in proteins usually lie in cavities, where the binding of a substrate typically serves as a mechanism for triggering some event, such as a chemical modification or conformational change. Structural information of protein surface regions enables detailed studies of the relationship of protein structure and function.

### Residue

(Informal:) A residue is a synonym for the side-chain of an amino acid or for the whole amino acid.

(Formal:) In a polypeptide chain the carboxyl group of amino acid  $n$  has formed a peptide bond, C–N, to the amino group of amino acid  $n+1$ . These repeating units are called residues.

### Protein Domain

PROTEIN MOLECULES ARE ORGANIZED IN A STRUCTURAL HIERARCHY

The primary structure is the arrangement of amino acids along a linear polypeptide chain. Two different proteins that have significant similarities in their primary structure are said to be homologous to each other.

Secondary structure occurs mainly as  $\alpha$  helices and  $\beta$  strands. These elements

usually arrange themselves in simple motifs (e.g. helix-loop-helix or hairpin). The main chain is arranged in secondary structure to neutralize its polar atoms through hydrogen bonds.

Several motifs usually combine to form compact structures, which are called domains.

The term tertiary structure is a common term both for the way motifs are arranged into domain structures and for the way a single polypeptide chain folds into one or several domains. Proteins that have only one chain are called monomeric proteins. But a fairly large number of proteins have a quaternary structure, which consists of several identical (same function) polypeptide chains (subunits) that associate into a multimeric molecule in a specific way. These subunits can function either independently of each other or cooperatively so that the function of one subunit is dependent on the functional state of other subunits. Other protein molecules are assembled from several different subunits with different function.

#### LARGE POLYPEPTIDE CHAINS FOLD INTO SEVERAL DOMAINS

The fundamental unit of tertiary structure is the domain. A domain is defined as a polypeptide chain or a part of a polypeptide chain that can fold independently into a stable tertiary structure. Domains are also units of function. Often, the different domains of a protein are associated with different functions (e.g. one domain for DNA binding and another one for dimerization with another protein). Proteins may comprise a single domain or as many as several dozen domains. There is no fundamental structural distinction between a domain and a subunit.

#### DOMAINS ARE BUILT FROM STRUCTURAL MOTIFS

Domains are formed by different combinations of secondary structure elements and motifs. The number of such combinations found in proteins is limited, and some combinations seem to be structurally favored. Thus similar domain structures frequently occur in different proteins with different functions and with completely different amino acid sequences.

Domains are classified into three main structural groups:  $\alpha$  structures, where the core is built up exclusively from  $\alpha$  helices;  $\beta$  structures, which comprise antiparallel  $\beta$  sheets; and *alpha/beta* structures, where combinations of  $\beta$ - $\alpha$ - $\beta$  motifs form a predominantly parallel  $\beta$  sheet surrounded by  $\alpha$  helices.

## Protein Signature

The genome sequencing centres are generating raw sequence data at an alarming rate, and the result is a need for automated sequence analysis methods. The automatic analysis of protein sequences is possible through the use of protein signatures, which are methods for identifying a domain or characteristic region of a protein family in a protein sequence.

Signatures are short amino acid sequences that are used to find homologous protein domains. The two short sequences of 15 and 9 amino acids shown (green) can be used to search large databases for a protein domain that is found in many proteins, the SH2 domain. Here, the first 50 amino acids of the SH2 domain of 100 amino acids is compared for the human and *Drosophila* Src protein. In the computer-generated sequence comparison (yellow row), exact matches between the human and *Drosophila* proteins are noted by the one-letter abbreviation for the amino acid; the positions with a similar but nonidentical amino acid are denoted by +, and nonmatches are blank. In this diagram, wherever one or both proteins contain an exact match to a position in the green sequences, both aligned sequences are colored red.

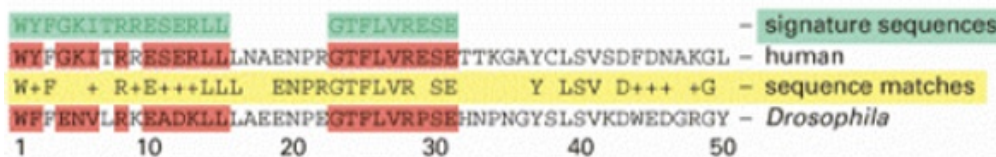


Figure 12.1: Using signatures to find homologous protein sequences.

## Sequence Homology Searches Can Identify Close Relatives

The present database of known protein sequences contains more than 500,000 entries, and it is growing very rapidly as more and more genomes are sequenced – revealing huge numbers of new genes that encode proteins. Powerful computer search programs are available that allow to compare each newly discovered protein with this entire database, looking for possible relatives. Homologous proteins are defined as those whose genes have evolved from a common ancestral gene, and these are identified by the discovery of statistically significant similarities in amino acid sequences.

With such a large number of proteins in the database, the search programs find many nonsignificant matches, resulting in a background noise level that makes it very difficult to pick out all but the closest relatives. Generally speaking, a 30% identity in the sequence of two proteins is needed to be certain that a match has been found. However, many short signature sequences (fingerprints) indicative of particular protein functions are known, and these are widely used to find more distant homologies.

These protein comparisons are important because related structures often imply related functions. Many years of experimentation can be saved by discovering that a new protein has an amino acid sequence homology with a protein of known function.

## Asymmetric Unit

When crystallographic structures are deposited in the PDB, the primary coordinate file generally contains one asymmetric unit - a concept that has applicability only to crystallography, but is important to understanding the process in obtaining the functional biological molecule.

An asymmetric unit is the smallest portion of a crystal structure to which crystallographic symmetry can be applied to generate one unit cell. The symmetry operations most commonly found in biological macromolecular structures are rotations, translations, and screws (combined rotation and translation). The unit cell is the smallest unit in a crystal that when translated in three dimensions makes up the entire crystal.

The figure below gives a simple example in two dimensions. Here, the asymmetric unit (green upward arrow) is rotated 180 degrees to produce a second copy (purple downward arrow). Together the two arrows comprise the unit cell. The unit cell is then translationally repeated in two directions to make up the entire crystal. The black oval in each unit cell represents the two-fold rotational symmetry axis that relates the green and purple arrows. In a real crystal, additional copies of the asymmetric unit may be required to make up the unit cell and the whole system would exist in three dimensions.

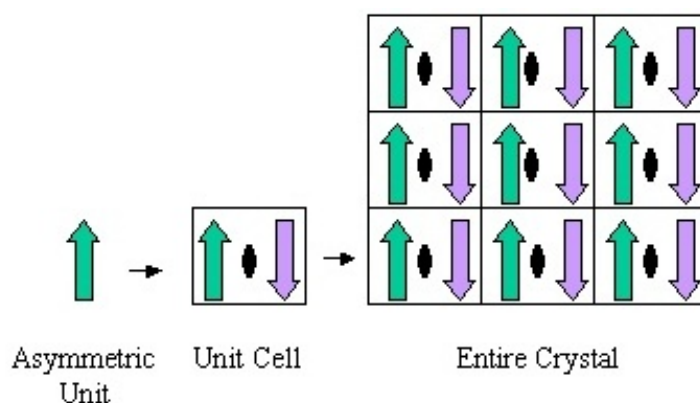


Figure 12.2: Asymmetric Unit: A simple example.

The asymmetric unit is used by the crystallographer to refine the structure against experimental data and does not necessarily represent a biologically functional molecule.



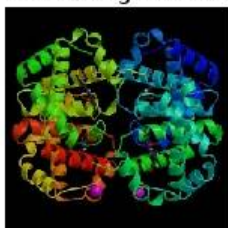
An asymmetric unit may contain:

- one biological molecule
- a portion of a biological molecule
- multiple biological molecules

The contents of the asymmetric unit depend on the molecule's position within the unit cell with respect to crystallographic symmetry elements and the level of structural similarities between multiple copies and structurally homologous portions of the molecule. Depending on crystallization conditions and local packing constraints, homologous copies of a protein, chain, or domain may take on slightly different conformations and cause the asymmetric unit to contain multiple structurally similar, but not exactly identical copies.

Hemoglobin, a molecule with four protein chains (two alpha-beta dimers), provides a good example of each of these cases:

#### One biological molecule



Entry [2hhb](#) contains **one** hemoglobin molecule (**4 chains**) in the asymmetric unit.

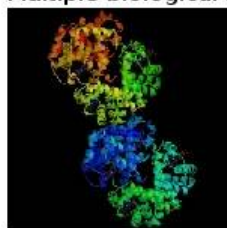
#### A portion of a biological molecule



Entry [1hho](#) contains **half** a hemoglobin molecule (**2 chains**) in the asymmetric unit. A crystallographic two-fold axis generates the 4 chains of the hemoglobin molecule.

The two homologous portions of the molecule are structurally similar enough that only one copy appears in the asymmetric unit.

#### Multiple biological molecules



Entry [1hv4](#) contains **two** hemoglobin molecules (**8 chains**) in the asymmetric unit.

Because of structural differences in the two hemoglobin molecules, both appear in the asymmetric unit.

Figure 12.3: Asymmetric Unit: possible contents of the asymmetric unit.

## Biological Unit = Biological Molecule

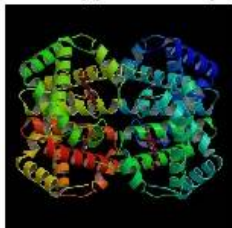
The biological molecule (also called a biological unit) is the macromolecule that has been shown to be or is believed to be functional. For example, the functional hemoglobin molecule has 4 chains. In each of the examples of hemoglobin mentioned above, the biological unit remains the same – 4 chains comprising one molecule of hemoglobin.

Depending on the asymmetric unit, spacegroup symmetry operations consisting of either rotations or translations must be performed in order to obtain the complete biological unit. However, if the asymmetric unit contains multiple biological molecules, then one copy may be selected. Thus a biological unit may be built from:

- one copy of the asymmetric unit
- multiple copies of the asymmetric unit
- a portion of the asymmetric unit

The hemoglobin example again demonstrates each of these cases:

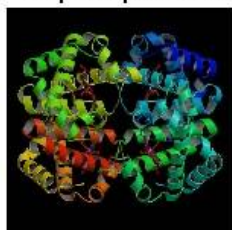
### One copy of the asymmetric unit



In entry [2hhb](#), the biological unit is **1** asymmetric unit.

No operation necessary.

### Multiple copies of the asymmetric unit



In entry [1hho](#), the biological unit is **2** asymmetric units.

Application of a crystallographic symmetry operation (a 180 rotation around a crystallographic two-fold axis) produces the complete biological unit.

### A portion of the asymmetric unit



In entry [1hv4](#), the biological unit is **half** the asymmetric unit.

PDB file contains **two** structurally similar, but not exactly identical copies of the biological unit.

Figure 12.4: The biological unit is build up from asymmetric units.

A biological unit is not always a multichain grouping. For example, the functional unit of dihydrofolate reductase is a monomer and thus the biological unit contains only one chain.

Occasionally, a molecule may appear multimeric in the crystal, but this has not been proven through other studies to be biologically relevant. For example, in the lysozyme structure presented in entry 104l , the asymmetric unit looks to be dimeric, but lysozyme is known to be functional as a monomer. Thus the biological unit is half of the asymmetric unit.

In certain cases, most notably viral capsids, the coordinate file may contain only part of the asymmetric unit. Here, the complete asymmetric unit can be generated by applying non-crystallographic symmetry operators to the coordinates. This complete asymmetric unit in turn may either form the biological unit (coat protein) or, in some complicated cases, only part of the biological unit. In the latter cases crystallographic symmetry operators may have to be applied to form the full biological unit (viral capsid). Non-crystallographic symmetry averaging is used experimentally to improve data quality.

For example, in the structure of the host range controlling region of feline parvovirus in entry 1p5y, non-crystallographic symmetry is used to create the the icosohedral viral capsid from sixty copies of the one protein chain contained in the coordinate file. The viral capsid is both the asymmetric unit and the biological unit.

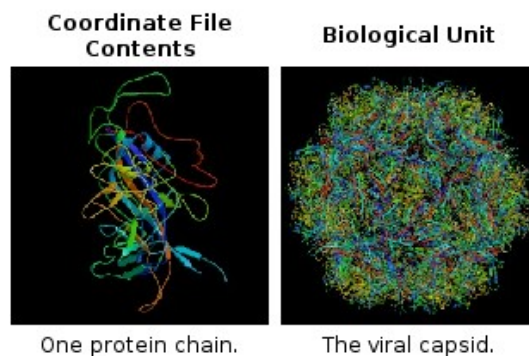


Figure 12.5: The .pdb file may contain only part of the asymmetric unit.

## Gene Ontology (GO) terms – biological process, molecular function, cellular component

GENE ONTOLOGY: TOOL FOR THE UNIFICATION OF BIOLOGY

The Gene Ontology project (GO) <http://www.geneontology.org/> is a dynamic controlled vocabulary defined in three ontology's, molecular function, biological process and cellular component. An ontology comprises a set of well-defined terms with well-defined relationships. The structure itself reflects the current representation of biological knowledge as well as serving as a guide for organizing new data. **Molecular function** is defined as the biochemical activity (including specific binding to ligands or structures) characteristic of a gene product. It describes only what is done without specifying where or when the event actually occurs. **Biological process** describes a phenomenon marked by changes that lead to a particular result, mediated by one or more gene products. Biological process refers to a biological objective to which the gene or gene product contributes. A process is accomplished via one or more ordered assemblies of molecular functions. **Cellular component** is the part of a cell of which a gene product is a component and where it is active; GO includes the extracellular environment of cells; a gene product may be a component of one or more parts of a cell.

Where once biochemists characterized proteins by their diverse activities and abundances, and geneticists characterized genes by the phenotypes of their mutations, all biologists now acknowledge that there is likely to be a single limited universe of genes and proteins, many of which are conserved in most or all living cells. This recognition has fuelled a grand unification of biology; the information about the shared genes and proteins contributes to our understanding of all the diverse organisms that share them. Knowledge of the biological role of such a shared protein in one organism can certainly illuminate, and often provide strong inference of, its role in other organisms. For the most part, the current systems of nomenclature for genes and their products remain divergent even when the experts appreciate the underlying similarities. Interoperability of genomic databases is limited by this lack of progress, and it is this major obstacle that the Gene Ontology (GO) Consortium was formed to address. A static hierarchical system, such as the Enzyme Commission (EC) hierarchy, although computationally tractable, was also likely to be inadequate to describe the role of a gene or a protein in biology in a manner that would be either intuitive or helpful for biologists.

AN EXAMPLE (MCM PROTEINS)

MCM proteins play a role in DNA metabolism: biological process ontology→DNA metabolism→DNA replication→DNA-dependent DNA replication→DNA unwinding, DNA initiation, pre-replicative complex formation and maintenance

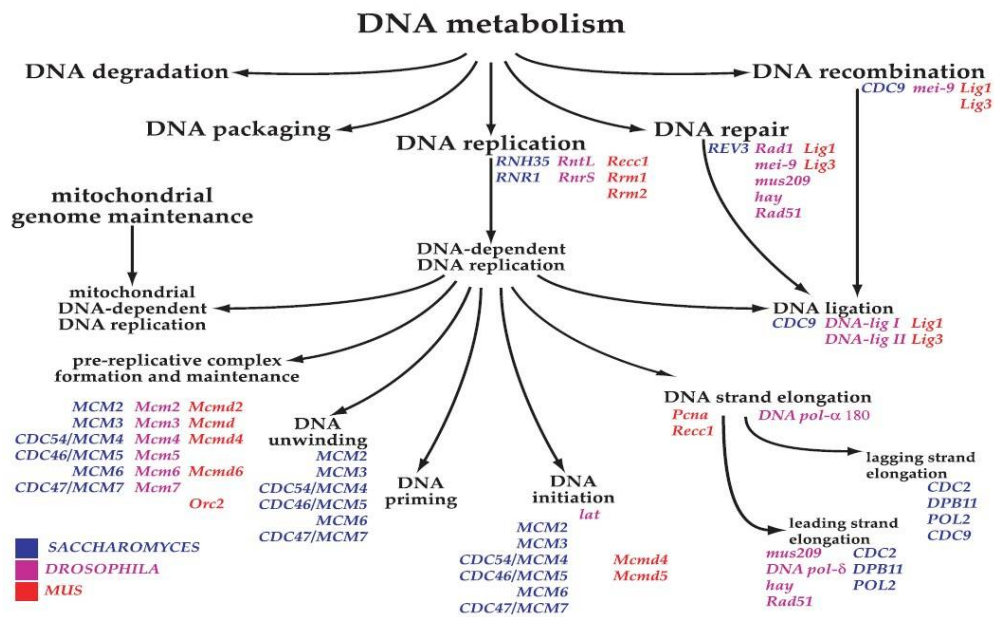


Figure 12.6: GO biological process: An example of GO annotation.

## Protein pocket

Pockets are empty concavities on a protein surface into which solvent can gain access, i.e., these concavities have mouth openings connecting their interior with the outside bulk solution. Pockets are defined as concave caverns with constrictions at the opening on the surface regions of proteins. Unlike voids, pockets allow easy access of water probes from the outside.

## Protein cavity/void

A cavity (or void) is an interior empty space that is not accessible to the solvent probe. It has no mouth openings to the outside bulk solution. Voids are defined as buried unfilled empty space inside proteins after removing all hetero atoms that are inaccessible to water molecules from outside.

## References

- 1 C. Branden, J. Tooze:  
*Introduction to Protein Structure*;  
Garland Publishing, 1999, Second Edition
- 2 B. Alberts, A. Johnson et al:  
*Molecular Biology of the Cell*;  
Garland Science, 2002, Fourth Edition
- 3 PDB online help